

OPTIMAL PROPOSAL DESIGN FOR RANDOM WALK TYPE METROPOLIS ALGORITHMS WITH GAUSSIAN RANDOM FIELD PRIORS

BY NATESH S. PILLAI, ANDREW M. STUART, ALEXANDRE H. THIÉRY

We study random walk based algorithms for posterior simulation in a large class of Bayesian nonparametric problems with Gaussian random field or Gaussian process priors. Our emphasis is both on developing practical guidelines for the design and implementation of efficient algorithms for these naturally high dimensional problems, and on the development of rigorous underpinning theory. We emphasize the principle of ‘optimal design’ of MCMC methods on the underlying infinite dimensional space in Bayesian nonparametric problems and show that this design principle leads to vastly improved, and novel, algorithms in finite dimensions.

To illustrate this idea we demonstrate that a small change in the *mean* of the proposal in a random walk algorithm can result in significant improvement in computational complexity as the dimension of the nonparametric approximation is refined. In particular we compare the computational complexity of this resulting new algorithm, pCN, with a standard random walk algorithm, RWM. RWM is optimized with respect to proposal *variance*, as suggested by the prevailing optimal scaling theory. We show that pCN exhibits an order of magnitude improvement in computational complexity over RWM, in terms of the dimension of the space.

Our methods of analysis rely on diffusion limits for the MCMC methods. This approach also enables us to develop a theory of simulated annealing for posterior simulation in an high (and infinite) dimensional settings, and shows how a noisy gradient descent algorithm can emerge, without explicitly computing the gradient, from certain carefully specified random walks; this results from the Metropolis-Hastings accept-reject mechanism. This theory extends results known in finite dimensions for finding local maxima using a gradient flow and is hence of independent interest.

1. Introduction.

1.1. *The Context.* Bayesian nonparametrics have witnessed a rapid growth in the last decade, both in the construction of novel prior distributions [DPP07, WCT11] and obtaining results on posterior consistency and sharp convergence rates, *e.g.*, [GVDV11, VDVZ08]. Nonparametric methods are increasingly popular in many applications due to their flexibility to model a wide variety of statistical problems when the parameter is naturally infinite dimensional. In such problems, the posterior distribution usually does not have a closed form solution and therefore one needs to resort to computational methods, such as the Markov Chain Monte Carlo (MCMC) algorithms, to get samples from the posterior distribution. In this regard, a wide array of fast computational algorithms have been developed for obtaining posterior draws [RC04, PR08, IZ00]. Here we focus on the optimal design of proposals for target measures arising in Bayesian nonparametrics with Gaussian random field priors. Gaussian random fields are ubiquitous in nonparametric modeling because of their remarkably rich sample properties, and the fact that they can be characterized entirely by the mean function and covariance operator [RW06]. In particular they arise in the Bayesian approach to inverse problems [Fit91, Stu10] and in the study of conditioned diffusion processes [HSV10].

Since the parameter space is infinite dimensional in these problems, practical implementation of MCMC involves discretizing the parameter space, resulting in a target measure in \mathbb{R}^N , with $N \gg 1$. It is well known that such discretization schemes can suffer from the curse of dimensionality: the efficiency of the algorithm decreases as the dimension N of the discretized space grows large. The central message of this paper is the following:

Optimal Proposal Design Principle. *Designing proposals which are well-defined on the infinite dimensional parameter space results in MCMC methods which do not suffer from the curse of dimensionality.*

In this work we will substantiate the above idea with a specific example consisting of two near-identical looking versions of the Random Walk Metropolis algorithm, namely the standard random walk (RWM) and a minor variant of it which we call the pCN algorithm (for preconditioned Crank-Nicolson, explained below), in the very general setting of Bayesian nonparametrics with Gaussian random field priors. The standard random walk proposal is not defined on the infinite dimensional parameter space and hence does not adhere to the optimal design principle. The prevailing tool currently available to practitioners for implementing the RWM is the optimal scaling result [RGG97], which suggests optimal tuning of the size of the *proposal variance*. On the other hand the pCN algorithm is defined on the infinite dimensional space, and the optimal proposal design principle applied to this algorithm suggests tuning the *proposal mean* instead of the *variance*. Even though the construction of the above two algorithms are nearly identical, and thus implementing pCN over RWM requires only a minor change in implementation, it turns out that the pCN algorithm enjoys an order of magnitude in efficiency gain as compared to the RWM. We will rigorously show that the optimal tuning of the size of the proposal variance for RWM as suggested by the optimal scaling results of [RGG97] has a negligible effect on computational complexity in comparison with the effect obtained from following our optimal proposal design principle, which tunes the *proposal mean* resulting in the pCN algorithm.

Before proceeding further, we would like to emphasize again that one of the key goals of our work is to convey the optimal design principle to practitioners. In particular, our exclusive focus on the RWM algorithm in this paper is made to demonstrate the efficacy of the optimal design principle in the simplest nontrivial example. We discuss this issue at the end of subsection 1.3 and in subsection 1.8.

1.2. The Target and The Algorithms. First we describe the class of models to which our main results are applicable. We denote the prior and posterior distributions respectively by π_0 and π and consider models in which both π_0 and π are measures on a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle, \|\cdot\|)$. Furthermore $\pi_0 = N(0, C)$ is assumed to be a Gaussian random field on \mathcal{H} and C is a covariance operator. The posterior π is given by the identity

$$\frac{d\pi}{d\pi_0}(x) = M_\Psi \exp(-\Psi(x)), \quad x \in \mathcal{H} \quad (1.1)$$

for a real valued functional Ψ (which denotes the negative log-likelihood) and M_Ψ a normalizing constant. Although the above formulation may appear quite abstract, we emphasize that this points to the wide-ranging applicability of our theory: the setting encompasses a large class of models arising in practice, including nonparametric regression and diffusion processes [HSV10, Stu10]. In Section 2.4 we discuss a few commonly used statistical models which belong to our framework.

Next we describe the class of MCMC algorithms which are our focus. Underlying the family of proposals we adopt is the assumption that it is straightforward to sample from the Gaussian

random field $N(0, C)$. The proposals we consider are hence of the form

$$(1.2) \quad y = ax + \sqrt{2\delta} \xi$$

where $\xi \sim N(0, C)$ is chosen independently of x , $\delta > 0$ sets the scale for the proposal variance and $a \in \mathbb{R}$ scales the proposal mean. Two instances of the algorithm are of particular interest: the standard random walk (RWM)

$$y = x + \sqrt{2\delta} \xi \quad (1.3)$$

and the preconditioned Crank-Nicolson walk (pCN)

$$y = (1 - 2\delta)^{\frac{1}{2}} x + \sqrt{2\delta} \xi. \quad (1.4)$$

The preconditioned random walk (pCN) proposal is introduced in [BRSV08] as the PIA algorithm, with the parameter choices $\theta = \frac{1}{2}$ and $\alpha = 0$; the nomenclature used here, namely the preconditioned Crank-Nicolson algorithm – pCN algorithm – is introduced and put in historical context in [CRSW11]. To implement the algorithms in practice a finite dimensional approximation of the target π is used. In our set-up, a natural way to perform the discretization is to consider the N -dimensional space spanned by the first N eigenfunctions of C , $\varphi_j, 1 \leq j \leq N$, project the target density into this basis and let the Markov chain evolve in \mathbb{R}^N . This is easy to implement as well, by setting in $\xi \sim N(0, C^N)$ in (1.2), where the diagonal elements of the covariance matrix C^N are the first N eigenvalues of C and the off-diagonal elements are 0.

Once this finite dimensional target is set-up, a key question for the practitioner is how to choose the free parameters a and δ in the proposal? In the context of the family of proposals (1.2), most of the practical guidance provided by statistical theory to date is focussed on the RWM proposal (1.3) and optimal tuning of the *proposal variance* δ . We will show that this has negligible effect on computational complexity when compared with choosing the *proposal mean scale* a . In particular the choice $a(\delta) = (1 - 2\delta)^{\frac{1}{2}}$ appearing in the pCN proposal (1.4) gives the optimal proposal design within the family of proposals (1.2).

1.3. Implications of Theory. Suppose we are interested in computing the expectation of test function $f : \mathcal{H} \rightarrow \mathbb{R}$ using MCMC methods based on the family of proposals (1.2). We assume that f is bounded with bounded (Fréchet) derivative. We let $\{x^{k,\delta}\}_{k \in \mathbb{Z}^+}$ denote the Markov chain resulting from using the proposal (1.2) on the finite dimensional approximation of π in \mathbb{R}^N and applying the Metropolis-Hastings accept-reject criterion.

Consider the following estimator for the expectation of f under the measure π given by

$$\hat{f}_{N,a,\delta} = \frac{1}{K} \sum_{k=0}^{K-1} f(x^{k,\delta}) \quad (1.5)$$

which is just the empirical mean of the Markov chain output. We let $C(N)$ denote the cost of evaluating one step of the Markov chain with proposal (1.2), including evaluation of the acceptance probability; this cost grows with dimension N , but is independent of a and δ . We ask the question: what is the computational cost, in terms of N , to ensure that $\hat{f}_{N,a,\delta}$ is within a tolerance level ϵ of $\mathbb{E}^\pi f$. We let \mathbb{E}^π denote expectation with respect to π for fixed starting point x^0 in the Markov chain, and let $\mathbb{E}^{\pi,*}$ denote expectation with respect to π for x^0 in stationarity, *i.e.*, $x^0 \sim \pi$. The following theorem is precisely stated and proved as Theorem 4.1.

Theorem 1: Optimal Proposal Mean Consider the MCMC method based on proposal (1.4) with optimally tuned mean $a = (1 - 2\delta)^{\frac{1}{2}}$. Then, for any $\epsilon > 0$ and any x^0 in the support of π , it is possible to choose K independent of N to ensure that

$$|\mathbb{E}^\pi \hat{f}_{N,a,\delta} - \mathbb{E}^\pi f| \leq \epsilon,$$

and the resulting computational cost grows with N as $C(N)$.

This theorem about computational cost, and the effect of scaling the *mean parameter*, will be a consequence of the theory we develop in this paper. For comparison it is instructive to see what optimal *variance scaling* suggested by the optimal scaling results for RWM [RGG97] tells us. This is the prevailing statistical theory which is adopted by practitioners to guide choice of proposals. The following assertion is stated more precisely and proved as Theorem 4.3, and is a direct consequence of recent theories which extend the work of [RGG97] to targets of the form (1.1) [MPS11]:

Theorem 2 (for Comparison): Optimal Proposal Variance Consider the MCMC method based on proposal (1.3) with optimally tuned variance $\delta = \ell \times N^{-1}$, to obtain acceptance probability 0.234, and $K = K_0 N^{-1}$. Then, for any $\epsilon > 0$ and $x^0 \sim \pi$, it is possible to choose K_0 independent of N so that

$$|\mathbb{E}^{\pi,*} \hat{f}_{N,a,\delta} - \mathbb{E}^\pi f| \leq \epsilon$$

and the resulting computational cost grows with N as $N \times C(N)$.

It is worth pausing to consider the implications of these two theorems for practical MCMC computations. In Bayesian inverse problems (especially in applied problems pertaining to ordinary and partial differential equations, see [Stu10] and the references there in) it is increasingly common to work with discretizations of size $N = \mathcal{O}(10^3)$ or even larger. The effect of focusing on the optimal proposal design principal that emerges from the theory in this paper results in an $\mathcal{O}(N)$ speed-up when compared with adopting the prevailing approach of tuning proposal variance. In practice this may mean the difference between running computations on a desktop rather than a multiprocessor computer, or between being able to evaluate desired expectations in reasonable computer time and not being able to do so, for example in on-line scenarios. Note also that the theory developed for the pCN variant of the RWM algorithm based on the optimal proposal design principle in this paper holds for *any initial condition* and is hence far more powerful than the theory based on tuning proposal variance, which works *only in stationarity*.

One of the celebrated aspects of the optimal proposal variance approach, widely cited and used by applied workers in many fields, is that practitioners should tune the average acceptance probability to be close to 0.234. There are even adaptive algorithms which are automated to obtain this acceptance probability. The two preceding theorems demonstrate that, whilst this choice of acceptance probability may have some benefits for RWM when applied to certain target measures, the practitioner would have far greater impact on complexity simply by adapting the proposal mean to obtain pCN, when sampling target measures of the form (1.1); furthermore, when doing so there is no universal optimal choice of acceptance probability.

These considerations should convince the reader that the stated **Optimal Design Principle** is worth pursuing further. Before proceeding to describe how we establish the theorems above, and describe the structure of the paper, we discuss the context for this design principle. We note first that, as stated, the principle is wide-ranging in applicability, but (deliberately) non-constructive. As mentioned above, here we demonstrate a particular application of the principle, with wide-ranging applicability to modifications of RWM proposals. The paper [CRSW11] shows how many

other new methods can be constructed by adhering to the design principle, and includes numerical demonstrations of their capabilities, and reference to other papers containing further numerical experiments. In particular, in addition to variants on standard RWM proposals, the paper shows how to modify Langevin, Hybrid (Hamiltonian) Monte-Carlo and Metropolis-within-Gibbs methods algorithms so as to adhere to the design principle in the context of the target measure (1.1). Finally we note that the optimal design principle has been known to improve the efficiency in many finite dimensional contexts. In the specific context of determining diffusion constants for partially observed diffusions, the design principle is adopted in [RS01] to break missing data/parameter dependencies and speed up algorithms. And in the context of spatial point processes, the principle is applied in [MW04]. The principle thus already has roots in the MCMC literature and the theory in this paper aims to both extend the understanding of this approach, and to provide theoretical tools to underpin it in the wide setting of Bayesian nonparametric problems.

1.4. *Diffusion Limits.* We establish the two theorems from the previous subsection by use of diffusion limits for RWM and pCN. A central role is played by the equation

$$(1.6) \quad dz(t) = -h\left(z(t) + C\nabla\Psi(z)\right) dt + \sqrt{2h}dW$$

where W is a Wiener process on \mathcal{H} with covariance operator C , and h is any positive scalar. This equation is π -reversible and ergodic, satisfying a law of large numbers for sample path averages [HSV07b]. Given the discrete time Markov chain x^k and the time increments $t_k = k\delta$, we define its piecewise linear interpolation to be:

$$z^\delta(t) = \frac{1}{\delta} (t - t_k) x^{k+1,\delta} + \frac{1}{\delta} (t_{k+1} - t) x^{k,\delta} \quad \text{for} \quad t_k \leq t < t_{k+1}. \quad (1.7)$$

Our key tool in proving the theorems of the last subsection will be the fact that z^δ converges weakly to z solving (1.6).

We pause to note that existence of a diffusion limit for the pCN algorithm is not entirely surprising. After all, most RWM algorithms evolve via local moves obtained by the discretization of an underlying diffusion and thus in the small noise limit will have a diffusive behavior. However it is note worthy that both RWM and pCN converge to the same diffusion, and that the relative efficiency of the two algorithms can be understood by examining the manner in which this limit is obtained. This is at the heart of demonstrating that the computational complexity of pCN is an order of magnitude smaller than that of RWM. And it is by virtue of conforming to the optimal design principle that we obtain this speed-up. The optimal design principle also applies to other algorithms, such as Hybrid Monte Carlo; but since such algorithms do not exhibit diffusive behaviour, other tools will be needed to study the optimal design principle in this context; this point is address in the discussion in subsection 1.8.

The idea of finding diffusion limits for MCMC methods was pioneered by Roberts and co-workers in [RGG97, RR98] for the RWM and MALA algorithms applied to i.i.d. targets; see [RR01] for an overview. Because the target measure for this work is an independent product it is possible to find decoupled scalar diffusions in each coordinate. The recent paper [MPS11] shows that similar limit theorems can be obtained for RWM applied to non-product measures of the form (1.1), by exploiting the Gaussian prior structure. The limiting diffusion now couples all coordinates and is given by the stochastic partial differential equation (SPDE) (1.6). In this paper we demonstrate that such limit theorems may also be obtained for pCN when applied to (1.1). The difference between the complexity bounds contained in the two theorems from subsection 1.3 may now be understood

heuristically as follows. By the ergodic theorem, the SPDE (1.6) needs to reach time $T = T(\epsilon)$ sufficiently large to obtain sample path averages which are $\mathcal{O}(\epsilon)$ close to the ergodic average. The limit theorem for RWM requires that $\delta = \mathcal{O}(N^{-1})$ and so $\mathcal{O}(N)$ steps are required to reach time T . In contrast the limit theorem for pCN is valid on spaces of all dimensions for the same δ sufficiently small; indeed the limit theorem holds on the infinite dimensional space \mathcal{H} . As a consequence the number of steps required to reach time T is $\mathcal{O}(1)$ with respect to dimension N for pCN.

1.5. *Insights from numerical analysis and optimal tuning of the proposal mean.* An important point to be noticed is that the diffusion limits for RWM algorithms are constructed with a “discretize then design a sampler” viewpoint: firstly, the problem (target distribution) is discretized, secondly, a standard Random Walk Metropolis algorithm applied and then the algorithm is tuned to achieve optimality. The advantage is that the algorithms need not be defined on the entire Hilbert space \mathcal{H} , but only on finite dimensional subspaces. In fact, the RWM is not well-defined on all of \mathcal{H} . One of our key observations in this paper is that, this also turns out to be the key disadvantage of the RWM algorithm in high dimensions. In contrast, the pCN algorithms adhere to the “design a sampler then discretize” principle (see [HPUU08], Chapter 3, for discussion of similar issues for optimization problems on function spaces), where the algorithms are defined on all of \mathcal{H} , and therefore, unlike RWM, do not suffer from the degeneracy due to discretization.

To further illustrate our viewpoint on optimal proposal design, let us give some intuition see why $a = (1 - 2\delta)^{1/2}$ is the right choice as suggested by the optimal design principle, behind the efficiency gain of pCN relative to all other proposals of the form (1.2). The key observation is that only the pCN proposal preserves the prior measure π_0 . To see this note that the Ornstein-Uhlenbeck (OU) process

$$\begin{aligned} dz &= -z dt + \sqrt{2\delta} dW \\ z_0 &= x, \end{aligned} \tag{1.8}$$

is π_0 -reversible and ergodic [DPZ96]; as for (1.6) here W is a Brownian motion in \mathcal{H} with covariance operator equal to C . If $t > 0$ then the exact solution of this equation has the form, for $\delta = \frac{1}{2}(1 - e^{-2t})$,

$$\begin{aligned} z(t) &= e^{-t}x + \sqrt{\left((1 - e^{-2t})\right)}\xi \\ &= (1 - 2\delta)^{\frac{1}{2}}x + \sqrt{2\delta}\xi, \end{aligned} \tag{1.9}$$

where $\xi \sim N(0, C)$. Thus the pCN proposal is an exact solution of this OU process.

The consequence of this observation is as follows: if $\Psi = 0$, the pCN algorithm started at stationarity has an acceptance probability of 1. For $\Psi \neq 0$ the acceptance probability has the form

$$\alpha^\delta(x, \xi) = 1 \wedge \exp(\Psi(x) - \Psi(y)) \tag{1.10}$$

(see Section 3 for more details). This acceptance probability is well-defined on \mathcal{H} and is $\mathcal{O}(1)$ for any $\delta \in (0, \frac{1}{2})$.

Now consider the proposal (1.2). It is straightforward to see that if $x \sim \pi_0$ then $y \sim N(0, (a^2 + 2\delta)C)$ and hence the prior is invariant if and only if $a^2 = 1 - 2\delta$. The acceptance probability now has the form

$$\alpha^\delta(x, \xi) = 1 \wedge \exp(I_{a,\delta}(x) - I_{a,\delta}(y)) \tag{1.11}$$

where

$$I_{a,\delta}(x) := \left(\frac{1}{2} + \frac{(a^2 - 1)}{4\delta} \right) \|C^{-\frac{1}{2}}x\|^2 + \Psi(x).$$

An important point to note here is that $\|C^{-\frac{1}{2}}x\|$ and $\|C^{-\frac{1}{2}}y\|$ are almost surely infinite with respect to π_0 and hence π . Hence, unless $a^2 = 1 - 2\delta$, the exponent in the acceptance probability approaches infinity as the dimension $N \rightarrow \infty$, for fixed δ . To control this effect for $a^2 \neq 1 - 2\delta$ requires $\delta \rightarrow 0$ as $N \rightarrow \infty$; this is at the heart of the theories of optimal proposal variance [RR01]. Since the number of steps to reach time T is inversely proportional to δ^{-1} in the diffusion limit scaling, this increases the computational complexity by a factor of $\delta^{-1}(N)$ when compared with the algorithm for $a^2 = 1 - 2\delta$ for which δ may be chosen independently of N . To be concrete we will confine attention in this paper to the cases $a = 1$, which is a standard RWM method, and $a^2 = 1 - 2\delta$ which gives the pCN method that we are advocating. However similar results could be derived for other values of a and would exhibit similar results to those for the RWM if $a = 1 + \mathcal{O}(\delta)$.

1.6. Simulated Annealing. We now describe a secondary motivation for study of the algorithms in this paper. There are many applications, including inverse problems in engineering and calculus of variations problems from materials science ([EHN96],[Dac89]) where it is of natural interest to find global or local minima of a functional

$$(1.12) \quad J(x) = \frac{1}{2} \|C^{-1/2}x\|^2 + \Psi(x) ,$$

where C is the self-adjoint, positive and trace-class linear operator described above, on the Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle, \|\cdot\|)$. Gradient flow or steepest descent is a natural approach to this problem, but in its basic form requires computation of the gradient $\nabla \Psi$ which, in some applications, may be an expensive or complex task. In addition, when multiple minima are present, it may be important to include noise within the algorithm in order to allow escape from local minima. The second goal of this paper is to show how a noisy gradient descent can emerge, *without explicitly computing the gradient*, from certain carefully specified random walks, when combined with a Metropolis-Hastings accept-reject mechanism [Tie98], with tunable noise level τ . In the finite state [KJV83, Čer85] or finite dimensional context [Gem85, GH86, HKS89] the idea of using random walks, with accept-reject, to perform global optimization is a well-known idea which goes by the name of simulated-annealing; see the review [BT93] for further references. The novelty of our work is that the theory is developed on an infinite dimensional Hilbert space, and the applications of the theory are particularly tailored towards practical problems arising in Bayesian nonparametrics. In this context, minimization of J given by (1.12) corresponds to finding the MAP estimator.

In finite dimensions the basic idea behind simulated annealing is built from Metropolis-Hastings methods which have an invariant measure with Lebesgue density proportional to $\exp(-\tau^{-1}J(x))$. By adapting the temperature $\tau \in (0, \infty)$ according to an appropriate cooling schedule it is possible to locate global minima of J . The essential challenge in transferring this idea to infinite dimensions is that there is no Lebesgue measure. However, our key observation in this paper is that, the non-existence of the Lebesgue measure can be circumvented by working with measures defined via their density with respect to a Gaussian measure like the posterior measure π defined in (1.1). To introduce the parameter τ , we modify our prior distribution π_0 and write

$$\pi_0^\tau = \mathcal{N}(0, \tau C), \tag{1.13}$$

so that the posterior distribution π^τ takes the form:

$$\frac{d\pi^\tau}{d\pi_0^\tau}(x) \propto \exp\left(-\frac{\Psi(x)}{\tau}\right). \quad (1.14)$$

Note that if \mathcal{H} is finite dimensional then π^τ has Lebesgue density proportional to $\exp(-\tau^{-1}J(x))$. It is also known that small ball probabilities are asymptotically maximized (in the small radius limit), under π^τ , on balls centred at minimizers of J [DS11]. To further incorporate the parameter τ , we modify the pCN proposal to be

$$y = (1 - 2\delta)^{\frac{1}{2}}x + \sqrt{2\delta\tau}\xi \quad (1.15)$$

where $\xi \sim N(0, C)$ and the “proposed move” (1.15) will be accepted or rejected with probability found from pointwise evaluation of Ψ given by,

$$\alpha^\delta(x, \xi) = 1 \wedge \exp\left(-\frac{1}{\tau}(\Psi(y) - \Psi(x))\right) \quad (1.16)$$

(see Section 3 for more details) resulting in a Markov chain $\{x^{k,\delta}\}_{k \in \mathbb{Z}^+}$.

A small generalization of the diffusion limit for pCN discussed in subsection 1.4 shows that, as δ goes to 0, the linear interpolation process $z^\delta(t)$ given by (1.7) now converges to the diffusion

$$(1.17) \quad dz(t) = -\left(z(t) + C\nabla\Psi(z)\right)dt + \sqrt{2\tau}dW$$

where W is a Wiener process on \mathcal{H} with covariance operator C . This equation is π^τ -reversible, ergodic and satisfies a law of large numbers with respect to the measure [HSV07b]. Since small ball probabilities under π^τ are maximized when centred at minimizers of J , the result thus shows that the algorithm will generate sequences which concentrate near minimizers of J . Varying τ according to a cooling schedule then results in a simulated annealing method on Hilbert space. Weak convergence results for the approximation of stochastic equations in infinite dimensions may be found in the numerical analysis literature. For the heat equation and variants see [Sha03, DP09, GKL09, KLL10], for dispersive and nondispersive wave problems see [Hau10, dBD06] and for delay equations see [BS05, BKMS08]. These papers rely on use of the Kolmogorov equation to establish weak convergence and do not typically deliver convergence on pathspace, but rather convergence of functionals at a given fixed time. In contrast our approach, which is much simpler, proves weak convergence on pathspace, and does not use the Kolmogorov equation; rather we use an invariance principle for Brownian motion in Hilbert space [Ber86], coupled with the preservation of weak convergence under continuous mappings. Our approach as it stands, does not deliver rates of weak convergence, but can be made more quantitative to obtain convergence rates. Since we are only interested in qualitative properties and their statistical applications, we do not venture in this direction.

Let us give a quick heuristic to see why the gradient flow emerges through the pointwise computation of Ψ and the accept-reject mechanism. Note that for $\delta \ll 1$, we see from (1.16) that

$$(1.18) \quad \alpha^\delta(x, \xi) \approx 1 \wedge \exp\left(-\sqrt{\frac{2\delta}{\tau}}\Psi(x)\xi\right).$$

This induces a bias towards accepting moves for which the the Gaussian random variable ξ , which is independent of x , aligns with the negative gradient of Ψ . Formalizing this heuristic is the content of Section 5.

Because the SDE (1.6) and the Markov chain are not started at stationarity, and because neither possess a smoothing property (they are only asymptotically strong Feller [HSV07a]), almost sure fine scale properties under its' invariant measure π^τ are not necessarily reflected at any finite time. For example if C is the covariance operator of Brownian motion or Brownian bridge then the quadratic variation of draws from the invariant measure, an almost sure quantity, is not reproduced at any finite time in (1.6) unless $z(0)$ has this quadratic variation; the almost sure property is approached only asymptotically as $t \rightarrow \infty$. This behaviour is reflected in the underlying Metropolis-Hastings Markov chain pCN which approximates (1.6), where the almost sure property is only reached asymptotically as $k \rightarrow \infty$. A third theorem that we highlight in this introduction (informally stated here and rigorously formulated and proved in Section 6) gives quantitative information about the rate at which the pCN algorithm approaches statistical equilibrium.

Theorem 3: Convergence of Almost Sure Quantities *The almost sure quantities such as the quadratic variation under pCN satisfy a limiting linear ordinary differential equation (ODE) with globally exponentially attractive steady state given by the value of the quantity under π^τ .*

We observe that the diffusion limit results obtained in this paper are entirely self-contained and the proof technique is analogous to the proof of diffusion limits of Markov chains in finite dimensions. We believe therefore that the methods of analysis that we introduce may be used to understand other MCMC algorithms and other nonparametric problems.

1.7. Structure of Paper. The paper is organized as follows. In section 2 we describe some notations used throughout the paper, discuss the required properties of Gaussian measures and Hilbert-space valued Brownian motions, and state our assumptions. In this section, we also exhibit a large class of statistical problems arising in practice which satisfy our assumptions. Section 3 contains a precise definition of the Markov chain $\{x^{k,\delta}\}_{k \in \mathbb{Z}^+}$, together with statement and proof of the weak convergence theorem to a diffusion that is the main theoretical result of the paper. Section 4 explains the consequences of this theory for computational complexity, stating precisely, and proving, the theorems on **Optimal Proposal Mean** and **Optimal Proposal Variance** from subsection 1.3. Section 5 contains proof of lemmas which underly the weak convergence theorem of section 3. In section 6 we state and prove the limit theorem for **Convergence of Almost Sure Quantities** such as quadratic variation; such results are often termed “fluid limits” in the applied probability literature. A simulation example illustrating the main result is presented in section 7. We conclude in section 8. Proofs of some technical lemmas are deferred to the Appendices A and B.

1.8. Discussion and Outlook. The basic design principle which we highlight in this paper is explained with reference to the random walk Metropolis algorithm and analysis of the computational complexity is undertaken via the use of diffusion limits. It is important that the reader appreciate the following two points. (i) The basic design principle is applicable far beyond the specific context of the random walk Metropolis methods; the paper [CRSW11] demonstrates the breadth of applicability of the idea, and contains a range of computational experiments to substantiate the benefits of adopting the new algorithms. (ii) The basic design principle can be understood theoretically from many different perspectives, and the recently developed spectral gap analyses of pCN and RWM [HSV11], using the 1-Wasserstein distance, provides a depth of understanding that is likely to transfer to many other situations where diffusion limits do not necessarily arise naturally; in particular, unlike the spectral gap results arising from the seminal work of Meyn and Tweedie [MT93], the work in [HSV11] is constructed to be well-adapted to infinite dimensional setting, and hence to analysis of

the basic design principle elucidated here. These links to recent emerging literature, and to other concluding observations, are made in section 8.

2. Preliminaries. In this section we define some notational conventions, Gaussian measure and Brownian motion in Hilbert space, and state our assumptions concerning the operator C and the functional Ψ .

2.1. Notation. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle, \|\cdot\|)$ denote a separable Hilbert space of real valued functions with the canonical norm derived from the inner-product. Let C be a positive, trace class operator on \mathcal{H} and $\{\varphi_j, \lambda_j^2\}_{j \geq 1}$ be the eigenfunctions and eigenvalues of C respectively, so that

$$C\varphi_j = \lambda_j^2 \varphi_j \quad \text{for} \quad j \in \mathbb{N}.$$

We assume a normalization under which $\{\varphi_j\}_{j \geq 1}$ forms a complete orthonormal basis in \mathcal{H} . For every $x \in \mathcal{H}$ we have the representation $x = \sum_j x_j \varphi_j$, where $x_j = \langle x, \varphi_j \rangle$. Using this notation, we define Sobolev-like spaces $\mathcal{H}^r, r \in \mathbb{R}$, with the inner-products and norms defined by

$$\langle x, y \rangle_r \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} j^{2r} x_j y_j \quad \text{and} \quad \|x\|_r^2 \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} j^{2r} x_j^2. \quad (2.1)$$

Notice that $\mathcal{H}^0 = \mathcal{H}$. Furthermore $\mathcal{H}^r \subset \mathcal{H} \subset \mathcal{H}^{-r}$ for any $r > 0$. The Hilbert-Schmidt norm $\|\cdot\|_C$ is defined as

$$\|x\|_C^2 = \sum_j \lambda_j^{-2} x_j^2.$$

For $r \in \mathbb{R}$, let $B_r : \mathcal{H} \mapsto \mathcal{H}$ denote the operator which is diagonal in the basis $\{\varphi_j\}_{j \geq 1}$ with diagonal entries j^{2r} , i.e.,

$$B_r \varphi_j = j^{2r} \varphi_j$$

so that $B_r^{\frac{1}{2}} \varphi_j = j^r \varphi_j$. The operator B_r lets us alternate between the Hilbert space \mathcal{H} and the interpolation spaces \mathcal{H}^r via the identities:

$$\langle x, y \rangle_r = \langle B_r^{\frac{1}{2}} x, B_r^{\frac{1}{2}} y \rangle \quad \text{and} \quad \|x\|_r^2 = \|B_r^{\frac{1}{2}} x\|^2. \quad (2.2)$$

Since $\|B_r^{-1/2} \varphi_k\|_r = \|\varphi_k\| = 1$, we deduce that $\{B_r^{-1/2} \varphi_k\}_{k \geq 1}$ forms an orthonormal basis for \mathcal{H}^r . For a positive, self-adjoint operator $D : \mathcal{H}^r \mapsto \mathcal{H}^r$, its trace in \mathcal{H}^r is defined as

$$\text{Trace}_{\mathcal{H}^r}(D) \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \langle B_r^{-\frac{1}{2}} \varphi_j, D B_r^{-\frac{1}{2}} \varphi_j \rangle_r.$$

Since $\text{Trace}_{\mathcal{H}^r}(D)$ does not depend on the orthonormal basis, the operator D is said to be trace class in \mathcal{H}^r if $\text{Trace}_{\mathcal{H}^r}(D) < \infty$ for some, and hence any, orthonormal basis of \mathcal{H}^r . Let $\otimes_{\mathcal{H}^r}$ denote the outer product operator in \mathcal{H}^r defined by

$$(x \otimes_{\mathcal{H}^r} y)z \stackrel{\text{def}}{=} \langle y, z \rangle_r x \quad \forall x, y, z \in \mathcal{H}^r. \quad (2.3)$$

For an operator $L : \mathcal{H}^r \mapsto \mathcal{H}^l$, we denote its operator norm by $\|\cdot\|_{\mathcal{L}(\mathcal{H}^r, \mathcal{H}^l)}$ defined by

$$\|L\|_{\mathcal{L}(\mathcal{H}^r, \mathcal{H}^l)} \stackrel{\text{def}}{=} \sup_{\|x\|_r=1} \|Lx\|_l.$$

For self-adjoint L and $r = l = 0$ this is, of course, the spectral radius of L .

Throughout we use the following notation.

- Two sequences $\{\alpha_n\}_{n \geq 0}$ and $\{\beta_n\}_{n \geq 0}$ satisfy $\alpha_n \lesssim \beta_n$ if there exists a constant $K > 0$ satisfying $\alpha_n \leq K\beta_n$ for all $n \geq 0$. The notations $\alpha_n \asymp \beta_n$ means that $\alpha_n \lesssim \beta_n$ and $\beta_n \lesssim \alpha_n$.
- Two sequences of real functions $\{f_n\}_{n \geq 0}$ and $\{g_n\}_{n \geq 0}$ defined on the same set Ω satisfy $f_n \lesssim g_n$ if there exists a constant $K > 0$ satisfying $f_n(x) \leq Kg_n(x)$ for all $n \geq 0$ and all $x \in \Omega$. The notations $f_n \asymp g_n$ means that $f_n \lesssim g_n$ and $g_n \lesssim f_n$.
- The notation $\mathbb{E}_x[f(x, \xi)]$ denotes expectation with variable x fixed, while the randomness present in ξ is averaged out.

2.2. Gaussian Measure on Hilbert Space. The following facts concerning Gaussian measure on Hilbert space, and Brownian motion in Hilbert space, may be found in [DPZ92]. Since C is self-adjoint, positive and trace-class we may associate with it a centred Gaussian measure π_0 on \mathcal{H} with covariance operator C , *i.e.*, $\pi_0 \stackrel{\text{def}}{=} \mathcal{N}(0, C)$. If $x \stackrel{\mathcal{D}}{\sim} \pi_0$ then we may write (Karhunen-Lo  ve)

$$x = \sum_{j=1}^{\infty} \lambda_j \rho_j \varphi_j \quad \text{with} \quad \rho_j \stackrel{\mathcal{D}}{\sim} \mathcal{N}(0, 1) \text{ i.i.d.} \quad (2.4)$$

If we let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space for the iid sequence $\{\varphi_j\}_{j \geq 1}$ then, since C is trace-class, the sum converges in $L^2(\Omega; \mathcal{H})$.

Since $\{B_r^{-1/2} \varphi_j\}_{j \geq 1}$ forms an orthonormal basis for \mathcal{H}^r , we may write (2.4) as

$$x = \sum_{j=1}^{\infty} (\lambda_j j^r) \rho_j (B_r^{-1/2} \varphi_j) \quad \text{with} \quad \rho_j \stackrel{\mathcal{D}}{\sim} \mathcal{N}(0, 1) \text{ i.i.d.} \quad (2.5)$$

Define

$$C_r = B_r^{1/2} C B_r^{1/2}. \quad (2.6)$$

If $\text{Trace}_{\mathcal{H}^r}(C_r) = \sum_{j=1}^{\infty} \lambda_j^2 j^{2r} < \infty$ for some $r > 0$ then the sum (2.5), and hence the sum (2.4), converges in $L^2(\Omega; \mathcal{H}^r)$ to a centred Gaussian random variable x with covariance operator C_r in \mathcal{H}^r , and C in \mathcal{H} . Then for any $u, v \in \mathcal{H}^r$, if $x \stackrel{\mathcal{D}}{\sim} \pi_0$, we have $\mathbb{E}[\langle x, u \rangle_r \langle x, v \rangle_r] = \langle u, C_r v \rangle_r$. Thus in what follows, we freely alternate between the Gaussian measures $\mathcal{N}(0, C)$ on \mathcal{H} and $\mathcal{N}(0, C_r)$ on \mathcal{H}^r . We note that

$$\mathbb{E}[\|x\|_r^2] = \sum_{j=1}^{\infty} \lambda_j^2 j^{2r} = \text{Trace}_{\mathcal{H}^r}(C_r).$$

Frequently in applications the functional Ψ arising in (1.12) may not be defined on all of \mathcal{H} , but only on a subspace $\mathcal{H}^s \subset \mathcal{H}$, for some exponent $s > 0$. Even though the Gaussian measure π_0 is defined on \mathcal{H} , depending on the decay of the eigenvalues of C , there exists an entire range of values of r such that $\text{Trace}_{\mathcal{H}^r}(C_r) < \infty$: in that case the measure π_0 has full support on \mathcal{H}^r , *i.e.*,

$\pi_0(\mathcal{H}^r) = 1$. Indeed, the condition $\text{Trace}_{\mathcal{H}^r}(C_r) < \infty$ also implies that for any $\tau > 0$ the measure π_0^τ has full support on \mathcal{H}^r . From now onwards we fix a distinguished exponent $s > 0$ and assume that $\Psi : \mathcal{H}^s \rightarrow \mathbb{R}$ and that $\text{Trace}_{\mathcal{H}^s}(C_s) < \infty$. Then $\pi_0 = N(0, C)$ on \mathcal{H} and $\pi(\mathcal{H}^s) = 1$. For ease of notations we introduce

$$\hat{\varphi}_j = B_s^{-\frac{1}{2}} \varphi_j \quad \text{for} \quad j \geq 1.$$

The family $\{\hat{\varphi}_j\}_{j \geq 1}$ forms an orthonormal basis for $(\mathcal{H}^s, \langle \cdot, \cdot \rangle_s)$. We may view the Gaussian measure $\pi_0 = N(0, C)$ on $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ as a Gaussian measure $N(0, C_s)$ on $(\mathcal{H}^s, \langle \cdot, \cdot \rangle_s)$.

A Brownian motion $\{W(t)\}_{t \geq 0}$ in \mathcal{H}^s with covariance operator $C_s : \mathcal{H}^s \rightarrow \mathcal{H}^s$ is a continuous Gaussian process with stationary increments satisfying $\mathbb{E}[\langle W(t), x \rangle_s \langle W(t), y \rangle_s] = t \langle x, C_s y \rangle_s$. For example, taking $\{\beta_j(t)\}_{j \geq 1}$ independent standard real Brownian motions, the process

$$W(t) = \sum_j (j^s \lambda_j) \beta_j(t) \hat{\varphi}_j \tag{2.7}$$

defines a Brownian motion in \mathcal{H}^s with covariance operator C_s ; equivalently, this same process $\{W(t)\}_{t \geq 0}$ can be described as a Brownian motion in \mathcal{H} with covariance operator equal to C since Equation (2.7) may also be expressed as $W(t) = \sum_{j=1}^\infty \lambda_j \beta_j(t) \varphi_j$.

2.3. Assumptions. In this section we describe the assumptions on the covariance operator C of the Gaussian measure $\pi_0 = N(0, C)$ and the functional Ψ . For each $x \in \mathcal{H}^s$ the derivative $\nabla \Psi(x)$ is an element of the dual $(\mathcal{H}^s)^*$ of \mathcal{H}^s , comprising the linear functionals on \mathcal{H}^s . However, we may identify $(\mathcal{H}^s)^* = \mathcal{H}^{-s}$ and view $\nabla \Psi(x)$ as an element of \mathcal{H}^{-s} for each $x \in \mathcal{H}^s$. With this identification, the following identity holds

$$\|\nabla \Psi(x)\|_{\mathcal{L}(\mathcal{H}^s, \mathbb{R})} = \|\nabla \Psi(x)\|_{-s}$$

and the second derivative $\partial^2 \Psi(x)$ can be identified with an element of $\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$. To avoid technicalities we assume that $\Psi(x)$ is quadratically bounded, with first derivative linearly bounded and second derivative globally bounded. Weaker assumptions could be dealt with by use of stopping time arguments.

ASSUMPTIONS 2.1. *The functional Ψ and the covariance operator C satisfy the following assumptions.*

A1. Decay of Eigenvalues λ_j^2 of C : *there exists a constant $\kappa > \frac{1}{2}$ such that*

$$\lambda_j \asymp j^{-\kappa}. \tag{2.8}$$

A2. Domain of Ψ : *there exists an exponent $s \in [0, \kappa - 1/2)$ such Ψ is defined on \mathcal{H}^s .*

A3. Size of Ψ : *the functional $\Psi : \mathcal{H}^s \rightarrow \mathbb{R}$ satisfies the growth conditions*

$$0 \leq \Psi(x) \lesssim 1 + \|x\|_s^2.$$

A4. Derivatives of Ψ : *The derivatives of Ψ satisfy*

$$\|\nabla \Psi(x)\|_{-s} \lesssim 1 + \|x\|_s \quad \text{and} \quad \|\partial^2 \Psi(x)\|_{\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})} \lesssim 1.$$

REMARK 2.2. *The condition $\kappa > \frac{1}{2}$ ensures that $\text{Trace}_{\mathcal{H}^r}(C_r) < \infty$ for any $r < \kappa - \frac{1}{2}$: this implies that $\pi_0^\tau(\mathcal{H}^r) = 1$ for any $\tau > 0$ and $r < \kappa - \frac{1}{2}$.*

REMARK 2.3. The functional $\Psi(x) = \frac{1}{2}\|x\|_s^2$ is defined on \mathcal{H}^s and its derivative at $x \in \mathcal{H}^s$ is given by $\nabla\Psi(x) = \sum_{j \geq 0} j^{2s} x_j \varphi_j \in \mathcal{H}^{-s}$ with $\|\nabla\Psi(x)\|_{-s} = \|x\|_s$. The second derivative $\partial^2\Psi(x) \in \mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$ is the linear operator that maps $u \in \mathcal{H}^s$ to $\sum_{j \geq 0} j^{2s} \langle u, \varphi_j \rangle \varphi_j \in \mathcal{H}^{-s}$: its norm satisfies $\|\partial^2\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})} = 1$ for any $x \in \mathcal{H}^s$.

The Assumptions 2.1 ensure that the functional Ψ behaves well in a sense made precise in the following lemma.

LEMMA 2.4. Let Assumptions 2.1 hold.

1. The function $d(x) \stackrel{\text{def}}{=} -(x + C\nabla\Psi(x))$ is globally Lipschitz on \mathcal{H}^s :

$$\|d(x) - d(y)\|_s \lesssim \|x - y\|_s \quad \forall x, y \in \mathcal{H}^s. \quad (2.9)$$

2. The second order remainder term in the Taylor expansion of Ψ satisfies

$$|\Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y - x \rangle| \lesssim \|y - x\|_s^2 \quad \forall x, y \in \mathcal{H}^s. \quad (2.10)$$

PROOF. See [MPS11]. □

2.4. *Statistical Models Satisfying Our Assumptions..* We demonstrate two classes of models which satisfy our assumptions. We emphasize, however, that our theoretical development has not been optimized with respect to the assumptions made. We could weaken the assumptions at the price of considerably more involved proofs. In broad terms we expect the ideas in this paper to apply whenever interest is focussed on sampling measures which possess a well-behaved density with respect to a Gaussian random field prior.

2.4.1. *Nonparametric Regression..* Consider estimating an unknown function x observed at discrete points on a compact set:

$$y_k = x(t_k) + \epsilon_k \quad (2.11)$$

where $t_k \in \mathcal{T} \subset \mathbb{R}^d$ and ϵ_k are i.i.d errors, which are mean 0 and have a smooth density $g(\cdot)$. The unknown function is assumed to be in the Hilbert space $\mathcal{H} = L_2(\mathcal{T})$, i.e., $x \in L_2(\mathcal{T})$. We assume a centered Gaussian random field prior on the function x , $\pi_0 = N(0, C)$ satisfying (2.8). Clearly, the (negative) log-likelihood is given by

$$\Psi(x) = - \sum_k \log g(y_k - x(t_k)).$$

By ensuring enough decay of eigenvalues, there are a large class of covariance operators C so that item 1 of Assumptions 2.1 is satisfied. Indeed, since $g(\cdot)$ has Gaussian or heavier tails, the function Ψ grows at most quadratically, satisfying item 2 of Assumptions 2.1. By virtue of $g(\cdot)$ belonging to an exponential family, the functional Ψ is smooth. Furthermore imposing the condition on the growth of the function x at infinity (which is related to the eigenvalues of the prior covariance C), the required bounds on the derivatives of Ψ can be obtained.

As mentioned above the theoretical properties of these regressions models are very well studied. In fact sharp rates of posterior concentration are known, and remarkable connections are established between the convergence rates of the posterior distribution and the small ball probabilities of the prior distribution [VDVVZ08, Cas08] which in turn is related to the eigenvalues of the covariance operator. Interesting results are also known from a minimax perspective [Zha00].

2.4.2. *Statistical Inference for Diffusions.* Consider the following stochastic differential equation:

$$dX_t = (AX_t - BB'\nabla V(X_t, \theta)) dt + B dW_t \quad (2.12)$$

where V is a drift function, $A, B \in \mathbb{R}^{d \times d}$ and $\theta \in \mathbb{R}^d$ is an unknown parameter. The statistical goal is to estimate θ from the discrete observations $X_{t_k} = x_k$, $t_k \in [0, T]$. SDE models of the above kind have received tremendous attention in recent years (see [BRSV08] and the references therein).

From a Bayesian point of view, a natural way to proceed (after carefully choosing the prior distribution for θ) is to obtain posterior samples via data augmented Gibbs sampling. This will involve two steps: sampling the conditional distribution of θ given the augmented (full) path $X_t, t \in [0, T]$ and sampling the conditional distribution of the augmented path $X_t, t \in [0, T]$ given θ , satisfying the constraint $X_{t_k} = x_k$. Sampling conditional diffusions (*i.e.*, diffusions X_t conditioned to satisfy a few values $X_{t_k} = x_k$) is a very challenging problem in general since their dynamics are quite intractable.

However for SDEs given in (2.12), there is a simple and efficient way to proceed. Let π denote the law of the diffusion bridge X_t conditioned to have $X_0 = x_0$ and $X_1 = x_1$ (more points can be easily dealt with using the Markovian property of the diffusion X_t). Notice that π is a probability measure on $\mathcal{H} = L_2[[0, 1], \mathbb{R}^d]$. If $V = 0$, then we obtain

$$dX_t = AX_t dt + B dW_t, \quad X_0 = x_0, X_1 = x_1 \quad (2.13)$$

which is a Gaussian process whose dynamics is more tractable. Let π_0 denote the law of the diffusion given in (2.13). Then it is known that ([BRSV08]) π is absolutely continuous with respect to π_0 on $\mathcal{H} = L_2[[0, 1], \mathbb{R}^d]$ with

$$\begin{aligned} \frac{d\pi}{d\pi_0}(X) &= \exp\{-\Psi(X)\} \\ \Psi(X) &= \langle 1, \Psi_1(X) \rangle, \quad \Psi_1(x) = |B^{-1}f(x)|^2 + \frac{1}{2} \operatorname{div} f(x) + f(x)'(BB')^{-1}Ax \\ f(x) &= -BB'\nabla V(X_t, \theta). \end{aligned} \quad (2.14)$$

Thus from (2.14) we see that the target measure π satisfies our formulation. Moreover, under mild regularity conditions, it can be shown that the functional Ψ in (2.14) satisfies Assumptions 2.1 [BRSV08].

There are many more practical applications including image processing, non-linear function estimation from partial differential equations which model physical phenomena and signal processing, where the target distributions satisfy our formulation and assumptions, see [CRSW11] for a detailed account.

3. Diffusion Theorem. This section contains a precise statement of the algorithm, statement of the main theorem showing that piecewise linear interpolant of the output of the algorithm converges weakly to a noisy gradient flow, and proof of the main theorem. The proof of various technical lemmas is deferred to section 5.

3.1. *pCN Algorithm.* We now define the Markov chain in \mathcal{H}^s which is reversible with respect to the measure π^τ given by Equation (1.14). This is the pCN algorithm with proposal (1.4). Let $x \in \mathcal{H}^s$ be the current position of the Markov chain. The proposal candidate y is given by (1.9), so that

$$y = (1 - 2\delta)^{\frac{1}{2}}x + \sqrt{2\delta\tau}\xi \quad \text{where} \quad \xi = N(0, C) \quad (3.1)$$

and $\delta \in (0, \frac{1}{2})$ is a small parameter which we will send to zero in order to obtain the noisy gradient flow. In Equation (3.1), the random variable ξ is chosen *independent* of x . As described in [BRSV08] (see also [CDS11, Stu10]), at temperature $\tau \in (0, \infty)$ the Metropolis-Hastings acceptance probability for the proposal y is given by

$$\alpha^\delta(x, \xi) = 1 \wedge \exp\left(-\frac{1}{\tau}(\Psi(y) - \Psi(x))\right). \quad (3.2)$$

The chain is then reversible with respect to π^τ . The Markov chain $x^\delta = \{x^{k,\delta}\}_{k \geq 0}$ can be written as

$$x^{k+1,\delta} = \gamma^{k,\delta} y^{k,\delta} + (1 - \gamma^{k,\delta}) x^{k,\delta} \quad \text{where} \quad y^{k,\delta} = (1 - 2\delta)^{\frac{1}{2}} x^{k,\delta} + \sqrt{2\delta\tau} \xi^k. \quad (3.3)$$

Here the ξ^k are iid Gaussian random variables $N(0, C)$ and the $\gamma^{k,\delta}$ are Bernoulli random variables which account for the accept-reject mechanism of the Metropolis-Hastings algorithm,

$$\gamma^{k,\delta} \stackrel{\text{def}}{=} \gamma^\delta(x^{k,\delta}, \xi^k) \stackrel{\mathcal{D}}{\sim} \text{Bernoulli}\left(\alpha^\delta(x^{k,\delta}, \xi^k)\right). \quad (3.4)$$

The function $\gamma^\delta(x, \xi)$ can be expressed as $\gamma^\delta(x, \xi) = \mathbb{I}_{\{U < \alpha^\delta(x, \xi)\}}$ where $U \stackrel{\mathcal{D}}{\sim} \text{Uniform}(0, 1)$ is independent from any other source of randomness. The next lemma will be repeatedly used in the sequel: it states that the size of the jump $y - x$ is of order $\sqrt{\delta}$.

LEMMA 3.1. *Under Assumptions 2.1 and for any integer $p \geq 1$ the following inequality*

$$\mathbb{E}_x[\|y - x\|_s^p]^{\frac{1}{p}} \lesssim \delta \|x\|_s + \sqrt{\delta} \lesssim \sqrt{\delta} (1 + \|x\|_s)$$

holds for any $\delta \in (0, \frac{1}{2})$.

PROOF. The definition of the proposal (3.1) shows that $\|y - x\|_s^p \lesssim \delta^p \|x\|_s^p + \delta^{\frac{p}{2}} \mathbb{E}[\|\xi\|_s^p]$. Fernique's theorem [DPZ92] shows that ξ has exponential moments and therefore $\mathbb{E}[\|\xi\|_s^p] < \infty$. This gives the conclusion. \square

For future use, we define the local mean acceptance probability at the current position x via the formula

$$\alpha^\delta(x) = \mathbb{E}_x[\alpha^\delta(x, \xi)]. \quad (3.5)$$

3.2. *Diffusion Limit Theorem.* Fix a time horizon $T > 0$ and a temperature $\tau \in (0, \infty)$. The piecewise linear interpolant z^δ of the Markov chain (3.3) is defined by Equation (1.7). The following is the main result of this article. Note that “weakly” refers to weak convergence of probability measures.

THEOREM 3.2. *Let Assumptions 2.1 hold. Let the Markov chain x^δ start at fixed position $x_* \in \mathcal{H}^s$. Then the sequence of processes z^δ converges weakly to z in $C([0, T]; \mathcal{H}^s)$, as $\delta \rightarrow 0$, where z solves the \mathcal{H}^s -valued stochastic differential equation*

$$\begin{aligned} dz &= -\left(z + C \nabla \Psi(z)\right) dt + \sqrt{2\tau} dW \\ z_0 &= x_* \end{aligned} \quad (3.6)$$

and W is a Brownian motion in \mathcal{H}^s with covariance operator equal to C_s .

For conceptual clarity, we derive Theorem 3.2 as a consequence of the general diffusion approximation Lemma 3.5. Consider a separable Hilbert space $(\mathcal{H}^s, \langle \cdot, \cdot \rangle_s)$ and a sequence of \mathcal{H}^s -valued Markov chains $x^\delta = \{x^{k,\delta}\}_{k \geq 0}$. The martingale-drift decomposition with time discretization δ of the Markov chain x^δ reads

$$\begin{aligned} x^{k+1,\delta} &= x^{k,\delta} + \mathbb{E}[x^{k+1,\delta} - x^{k,\delta} | x^{k,\delta}] + \left(x^{k+1,\delta} - x^{k,\delta} - \mathbb{E}[x^{k+1,\delta} - x^{k,\delta} | x^{k,\delta}] \right) \\ &= x^{k,\delta} + d^\delta(x^{k,\delta}) \delta + \sqrt{2\tau\delta} \Gamma^\delta(x^{k,\delta}, \xi^k) \end{aligned} \quad (3.7)$$

where the approximate drift d^δ and volatility term $\Gamma^\delta(x, \xi^k)$ are given by

$$\begin{aligned} d^\delta(x) &= \delta^{-1} \mathbb{E}[x^{k+1,\delta} - x^{k,\delta} | x^{k,\delta} = x] \\ \Gamma^\delta(x, \xi^k) &= (2\tau\delta)^{-1/2} \left(x^{k+1,\delta} - x^{k,\delta} - \mathbb{E}[x^{k+1,\delta} - x^{k,\delta} | x^{k,\delta} = x] \right). \end{aligned} \quad (3.8)$$

Notice that $\{\Gamma^{k,\delta}\}_{k \geq 0}$, with $\Gamma^{k,\delta} \stackrel{\text{def}}{=} \Gamma^\delta(x^{k,\delta}, \xi^k)$, is a martingale difference array in the sense that $M^{k,\delta} = \sum_{j=0}^k \Gamma^{j,\delta}$ is a martingale adapted to the natural filtration $\mathcal{F}^\delta = \{\mathcal{F}^{k,\delta}\}_{k \geq 0}$ of the Markov chain x^δ . The parameter δ represents a time increment. We define the piecewise linear rescaled noise process by

$$W^\delta(t) = \sqrt{\delta} \sum_{j=0}^k \Gamma^{j,N} + \frac{t - t_k}{\sqrt{\delta}} \Gamma^{k+1,N} \quad \text{for} \quad t_k \leq t < t_{k+1}. \quad (3.9)$$

We now show that, as $\delta \rightarrow 0$, if the sequence of approximate drift functions $d^\delta(\cdot)$ converges in the appropriate norm to a limiting drift $d(\cdot)$ and the sequence of rescaled noise process W^δ converges to a Brownian motion then the sequence of piecewise linear interpolants z^δ defined by Equation (1.7) converges weakly to a diffusion process in \mathcal{H}^s . In order to state the general diffusion approximation Lemma 3.5, we introduce the following:

CONDITIONS 3.3. *There exists an integer $p \geq 1$ such that the sequence of Markov chains $x^\delta = \{x^{k,\delta}\}_{k \geq 0}$ satisfies*

1. **Convergence of the drift:** *there exists a globally Lipschitz function $d : \mathcal{H}^s \rightarrow \mathcal{H}^s$ such that*

$$\|d^\delta(x) - d(x)\|_s \lesssim \delta \cdot (1 + \|x\|_s^p) \quad (3.10)$$

2. **Invariance principle:** *as δ tends to zero the sequence of processes $\{W^\delta\}_{\delta \in (0, \frac{1}{2})}$ defined by Equation (3.9) converges weakly in $C([0, T], \mathcal{H}^s)$ to a Brownian motion W in \mathcal{H}^s with covariance operator C_s .*

3. **A priori bound:** *the following bound holds*

$$\sup_{\delta \in (0, \frac{1}{2})} \left\{ \delta \cdot \mathbb{E} \left[\sum_{k \leq T/\delta} \|x^{k,\delta}\|_s^p \right] \right\} < \infty. \quad (3.11)$$

REMARK 3.4. *The a-priori bound (3.11) can equivalently be stated as $\sup_{\delta \in (0, \frac{1}{2})} \left\{ \mathbb{E} \left[\int_0^T \|z^\delta(u)\|_s^p du \right] \right\} < \infty$.*

It is now proved that Conditions 3.3 are sufficient to obtain a diffusion approximation for the sequence of rescaled processes z^δ defined by equation (1.7), as δ tends to zero.

LEMMA 3.5. (General Diffusion Approximation for Markov chains)

Consider a separable Hilbert space $(\mathcal{H}^s, \langle \cdot, \cdot \rangle_s)$ and a sequence of \mathcal{H}^s -valued Markov chains $x^\delta = \{x^{k,\delta}\}_{k \geq 0}$ starting at a fixed position $x_* \in \mathcal{H}^s$,

$$x^{0,\delta} = x_* \quad \forall \delta \in (0; 1).$$

Suppose that the drift-martingale decompositions (3.7) of x^δ satisfy Conditions 3.3. Then the sequence of rescaled interpolants $z^\delta \in C([0, T], \mathcal{H}^s)$ defined by equation (1.7) converges weakly in $C([0, T], \mathcal{H}^s)$ to $z \in C([0, T], \mathcal{H}^s)$ given by the stochastic differential equation

$$\begin{aligned} dz &= d(z) dt + \sqrt{2\tau} dW \\ z_0 &= x_* \end{aligned} \tag{3.12}$$

where W is a Brownian motion in \mathcal{H}^s with covariance C_s .

PROOF. For the sake of clarity, the proof of Lemma 3.5 is divided into several steps.

- **Integral equation representation.**

Notice that solutions of the \mathcal{H}^s -valued SDE (3.12) are nothing else than solutions of the following integral equation,

$$z(t) = x_* + \int_0^t d(z(u)) du + \sqrt{2\tau} W(t) \quad \forall t \in (0, T), \tag{3.13}$$

where W is a Brownian motion in \mathcal{H}^s with covariance operator equal to C_s . We thus introduce the Itô map $\Theta : C([0, T], \mathcal{H}^s) \rightarrow C([0, T], \mathcal{H}^s)$ that sends a function $W \in C([0, T], \mathcal{H}^s)$ to the unique solution of the integral equation (3.13): solution of (3.12) can be represented as $\Theta(W)$ where W is an \mathcal{H}^s -valued Brownian motion with covariance C_s . As is described below, the function Θ is continuous if $C([0, T], \mathcal{H}^s)$ is topologized by the uniform norm $\|w\|_{C([0, T], \mathcal{H}^s)} \stackrel{\text{def}}{=} \sup\{\|w(t)\|_s : t \in (0, T)\}$. It is crucial to notice that the rescaled process z^δ , defined in Equation (1.7), satisfies

$$z^\delta = \Theta(\widehat{W}^\delta) \quad \text{where} \quad \widehat{W}^\delta(t) := W^\delta(t) + \frac{1}{\sqrt{2\tau}} \int_0^t [d^\delta(\bar{z}^\delta(u)) - d(z^\delta(u))] du. \tag{3.14}$$

In Equation (3.14), the quantity d^δ is the approximate drift defined in Equation (3.8) and \bar{z}^δ is the rescaled piecewise constant interpolant of $\{x^{k,\delta}\}_{k \geq 0}$ defined as

$$\bar{z}^\delta(t) = x^{k,\delta} \quad \text{for} \quad t_k \leq t < t_{k+1}. \tag{3.15}$$

The proof follows from a continuous mapping argument (see below) once it is proven that \widehat{W}^δ converges weakly in $C([0, T], \mathcal{H}^s)$ to W .

- **The Itô map Θ is continuous**

It can be proved that Θ is continuous as a mapping from $(C([0, T], \mathcal{H}^s), \|\cdot\|_{C([0, T], \mathcal{H}^s)})$ to itself. The usual Picard's iteration proof of the Cauchy-Lipschitz theorem of ODEs may be employed: see [MPS11].

- **The sequence of processes \widehat{W}^δ converges weakly to W**

The process $\widehat{W}^\delta(t)$ is defined by $\widehat{W}^\delta(t) = W^\delta(t) + \frac{1}{\sqrt{2\tau}} \int_0^t [d^\delta(\bar{z}^\delta(u)) - d(z^\delta(u))] du$ and Conditions 3.3 state that W^δ converges weakly to W in $C([0, T], \mathcal{H}^s)$. Consequently, to prove that $\widehat{W}^\delta(t)$ converges weakly to W in $C([0, T], \mathcal{H}^s)$, it suffices to verify that the sequences of processes

$$(\omega, t) \mapsto \int_0^t [d^\delta(\bar{z}^\delta(u)) - d(z^\delta(u))] du \quad (3.16)$$

converges to 0 in probability with respect to the supremum norm in $C([0, T], \mathcal{H}^s)$. By Markov's inequality, it is enough to check that

$$\lim_{\delta \rightarrow 0} \mathbb{E} \left[\int_0^T \|d^\delta(\bar{z}^\delta(u)) - d(z^\delta(u))\|_s du \right] = 0.$$

Conditions 3.3 states that there exists an integer $p \geq 1$ such that $\|d^\delta(x) - d(x)\| \lesssim \delta \cdot (1 + \|x\|_s^p)$ so that for any $t_k \leq u < t_{k+1}$ we have

$$\left\| d^\delta(\bar{z}^\delta(u)) - d(\bar{z}^\delta(u)) \right\|_s \lesssim \delta (1 + \|\bar{z}^\delta(u)\|_s^p) = \delta (1 + \|x^{k,\delta}\|_s^p). \quad (3.17)$$

Conditions 3.3 states that $d(\cdot)$ is globally Lipschitz on \mathcal{H}^s . Therefore, Lemma 3.1 shows that

$$\mathbb{E} \|d(\bar{z}^\delta(u)) - d(z^\delta(u))\|_s \lesssim \mathbb{E} \|x^{k+1,\delta} - x^{k,\delta}\|_s \lesssim \delta^{\frac{1}{2}} (1 + \|x^{k,\delta}\|_s). \quad (3.18)$$

From estimates (3.17) and (3.18) it follows that $\|d^\delta(\bar{z}^\delta(u)) - d(z^\delta(u))\|_s \lesssim \delta^{\frac{1}{2}} (1 + \|x^{k,\delta}\|_s^p)$. Consequently

$$\mathbb{E} \left[\int_0^T \|d^\delta(\bar{z}^\delta(u)) - d(z^\delta(u))\|_s du \right] \lesssim \delta^{\frac{3}{2}} \sum_{k\delta < T} \mathbb{E} [1 + \|x^{k,\delta}\|_s^p]. \quad (3.19)$$

The a-priori bound of Conditions 3.3 shows that this last quantity converges to 0 as δ converges to zero, which finishes the proof of Equation (3.16). This concludes the proof of $\widehat{W}^\delta(t) \Rightarrow W$.

- **Continuous mapping argument.**

It has been proved that Θ is continuous as a mapping from $(C([0, T], \mathcal{H}^s), \|\cdot\|_{C([0, T], \mathcal{H}^s)})$ to itself. The solutions of the \mathcal{H}^s -valued SDE (3.12) can be expressed as $\Theta(W)$ while the rescaled continuous interpolate z^δ also reads $z^\delta = \Theta(\widehat{W}^\delta)$. Since \widehat{W}^δ converges weakly in $(C([0, T], \mathcal{H}^s), \|\cdot\|_{C([0, T], \mathcal{H}^s)})$ to W as δ tends to 0, the continuous mapping theorem ensures that z^δ converges weakly in $(C([0, T], \mathcal{H}^s), \|\cdot\|_{C([0, T], \mathcal{H}^s)})$ to the solution $\Theta(W)$ of the \mathcal{H}^s -valued SDE (3.12). This ends the proof of Lemma 3.5. □

In order to establish Theorem 3.2 as a consequence of the general diffusion approximation Lemma 3.5, it suffices to verify that if Assumptions 2.1 hold then Conditions 3.3 are satisfied by the Markov chain x^δ defined in section 3.1. In section 5.2 we prove the following quantitative version of the approximation $d^\delta \approx d$ where $d(x) = -(x + C\nabla\Psi(x))$:

LEMMA 3.6. (Drift estimate)

Let Assumptions 2.1 hold and let $p \geq 1$ be an integer. Then the following estimate is satisfied,

$$\|d^\delta(x) - d(x)\|_s^p \lesssim \delta^{\frac{p}{2}}(1 + \|x\|_s^{2p}). \quad (3.20)$$

Moreover, the approximate drift d^δ is linearly bounded in the sense that

$$\|d^\delta(x)\|_s \lesssim 1 + \|x\|_s. \quad (3.21)$$

It follows from Lemma (3.6) that Equation (3.10) of Conditions 3.3 is satisfied as soon as Assumptions 2.1 hold. The invariance principle of Conditions 3.3 follows from the next lemma. It is proved in section 5.5.

LEMMA 3.7. (Invariance Principle)

Let Assumptions 2.1 hold. Then the rescaled noise process $W^\delta(t)$ defined in equation (3.9) satisfies

$$W^\delta \implies W$$

where \implies denotes weak convergence in $C([0, T]; \mathcal{H}^s)$, and W is a \mathcal{H}^s -valued Brownian motion with covariance operator C_s .

In section 5.4 it is proved that the following a priori bound is satisfied,

LEMMA 3.8. (A priori bound)

Consider a fixed time horizon $T > 0$ and an integer $p \geq 1$. Under Assumptions 2.1 the following bound holds,

$$\sup \left\{ \delta \cdot \mathbb{E} \left[\sum_{k\delta \leq T} \|x^{k,\delta}\|_s^p \right] : \delta \in (0, \frac{1}{2}) \right\} < \infty. \quad (3.22)$$

In conclusion, Lemmas 3.6 and 3.7 and 3.8 together show that Conditions 3.3 are consequences of Assumptions 2.1. Therefore, under Assumptions 2.1, the general diffusion approximation Lemma 3.5 can be applied: this concludes the proof of Theorem 3.2.

4. Implications for Computational Complexity. In this section we state and prove precise versions of Theorems 1 and 2 from the subsection 1.3 in the introduction. Recall that the target measure π is defined by (1.1):

$$\frac{d\pi}{d\pi_0}(x) = M_\Psi \exp(-\Psi(x)).$$

and our goal is to quantify the computational complexity of the RWM and pCN algorithms designed to approximate expectations of the form $\int f d\pi$ for suitable test functions $f : \mathcal{H} \mapsto \mathbb{R}$. In practice a natural way to discretize the RWM and pCN algorithms is to sample the target measure π^N , the measure obtained by projecting π onto the first N eigenfunctions of C . To define this recall that

$$C\varphi_j = \lambda_j^2 \varphi_j, \quad j \in \mathbb{N}.$$

Define by P^N the orthogonal projection in \mathcal{H} onto the span of $\{\varphi_j\}_{j=1}^N$ and $\pi_0^N = N(0, P^N C P^N)$. Then set

$$\frac{d\pi^N}{d\pi_0^N}(x) = M_{\Psi, N} \exp(-\Psi(P^N x)).$$

The resulting Markov chain to sample π^N on \mathcal{H} may be implemented on \mathbb{R}^N . We initialize at $x^{0,\delta} = P^N x_*$, $x_* \in \mathcal{H}$. We let \mathbb{E}^{π^N} denote expectation with respect to π^N for fixed starting point $x^{0,\delta}$ in the Markov chain, and let $\mathbb{E}^{\pi^N,*}$ denote expectation with respect to π^N for $x_* \sim \pi$ or equivalently $x^{0,\delta} \sim \pi^N$.

We may now prove Theorem 1, concerning the pCN method. This follows from an application of Theorem 3.2 on $P^N \mathcal{H}$ with $\tau = 1$. Recall that $C(N)$ is the cost of implementing a single proposal of RWM, including evaluation of the acceptance probability; and that (1.5) defines $\hat{f}_{N,a,\delta}$ the estimator of the expectation of f using the Markov chain pCN or RWM for $a = (1 - 2\delta)^{1/2}$ and $a = 1$ respectively.

THEOREM 4.1. *Let Assumptions 2.1 hold and let $f : \mathcal{H}^s \rightarrow \mathbb{R}$ be bounded with bounded Fréchet derivative. Let $a = (1 - 2\delta)^{1/2}$. Then for any fixed x_* in the support of π and any $\epsilon > 0$, there is $K \in \mathbb{N}$ and $\delta_c \in (0, \frac{1}{2})$ both independent of N and $N_c \in \mathbb{N}$ such that, for $\delta < \delta_c$ and $N > N_c$*

$$|\mathbb{E}^{\pi^N} \hat{f}_{N,a,\delta} - \mathbb{E}^\pi f| \leq \epsilon.$$

Thus the computational cost grows to achieve this error grows with N as $C(N)$.

PROOF. We first observe that the limit Theorem 3.2 holds on $P^N \mathcal{H}$ and the corresponding limiting SPDE is given by (1.17) with $\Psi = \Psi(P^N \cdot)$ and $C \mapsto C^N = P^N C P^N$. This follows from inspection of the proof, using the fact that the theorem is proved on the infinite dimensional space \mathcal{H} and the observation that the functional $\Psi(P^N \cdot)$ satisfies all of the Assumptions 2.1 with constants independent of N . We denote the limiting solution by $z^{(N)}$.

We then use the following decomposition of the error:

$$\begin{aligned} |\mathbb{E}^{\pi^N} \hat{f}_{N,a,\delta} - \mathbb{E}^\pi f| &\leq |\mathbb{E}^{\pi^N} \hat{f}_{N,a,\delta} - \mathbb{E}^{\pi^N} f| + |\mathbb{E}^{\pi^N} f - \mathbb{E}^\pi f| \\ &\leq \left| \mathbb{E}^{\pi^N} \hat{f}_{N,a,\delta} - \frac{1}{T} \int_0^T \mathbb{E}^{\pi^N} f(z^{(N)}(s)) ds \right| + \left| \frac{1}{T} \int_0^T \mathbb{E}^{\pi^N} f(z^{(N)}(s)) ds - \mathbb{E}^{\pi^N} f \right| \\ &\quad + |\mathbb{E}^{\pi^N} f - \mathbb{E}^\pi f|. \end{aligned}$$

First choose N sufficiently large so that the third item is less than $\epsilon/3$. The existence of such an N follows from Theorem 4.6 of [Stu10] and the Lipschitz continuity of f . The second item may be made less than $\epsilon/3$ by the properties of the SPDE (1.17) and the ergodic Theorem 4.11 from [HSV07b], again for some $T = T(\epsilon)$ independent of N . Without loss of generality we may choose T so that it is an integer multiple of δ . With this choice we consider the first item on the right-hand side and note that it may be written as

$$\begin{aligned} \left| \mathbb{E}^{\pi^N} \hat{f}_{N,a,\delta} - \frac{1}{T} \int_0^T \mathbb{E}^{\pi^N} f(z^{(N)}(s)) ds \right| &= \left| \frac{1}{T} \int_0^T \left(\mathbb{E}^{\pi^N} f(\bar{z}^\delta) - \mathbb{E}^{\pi^N} f(z^{(N)}(s)) \right) ds \right| \\ &\leq \left| \frac{1}{T} \int_0^T \left(\mathbb{E}^{\pi^N} f(\bar{z}^\delta) - \mathbb{E}^{\pi^N} f(z^\delta(s)) \right) ds \right| \\ &\quad + \left| \frac{1}{T} \int_0^T \left(\mathbb{E}^{\pi^N} f(z^\delta) - \mathbb{E}^{\pi^N} f(z^{(N)}(s)) \right) ds \right|. \end{aligned}$$

Now the first integral tends to zero by application of the estimate (3.18) with f replacing d , noting that f is Lipschitz; and the second integral tends to zero by the weak convergence given in Theorem 3.2 on $P^N \mathcal{H}$ with $\tau = 1$. Both these require choosing δ sufficiently small, independently of N ; it is then possible to ensure that the third term in the preceding bound is also less than $\epsilon/3$ and the theorem is proved. \square

To prove Theorem 2 we use the following theorem from [MPS11]:

THEOREM 4.2. *Let $\{x^{k,\delta}\}$ be the Markov chain on \mathbb{R}^N corresponding to the RWM proposal applied to sample π^N and z^δ be as defined in (1.7). Then the following hold:*

1. *If the chain is started in stationarity for the “optimal scale” $\delta = \ell/N$ then the average acceptance probability converges to a number $\alpha(\ell) \in (0, 1)$ and the linear interpolation process z^δ converges weakly to the diffusion (1.6) on the space of paths, $C([0, T], \mathcal{H})$ with $z(0) = x$ and W a Wiener process on \mathcal{H} with covariance operator C , for some $h(\ell) \in (0, \infty)$.*
2. *The speed factor $h(\ell)$ is maximized over the parameter ℓ if the average acceptance probability is chosen to be 0.234 to three decimal places.*

Now we precisely state and prove our claim about the complexity of the RWM:

THEOREM 4.3. *Let Assumptions 2.1 hold and let $f : \mathcal{H} \rightarrow \mathbb{R}$ be bounded with bounded Fréchet derivative. Set $a = 1$ and $\delta = \hat{\ell}/N$ where $\hat{\ell}$ is the ‘optimal value’ from Theorem 4.2. Then for $x_0 \sim \pi^N$, any $\epsilon > 0$ and $K = K_0 N$ there is $K_0 \in \mathbb{N}$ independent of N and $N_c \in \mathbb{N}$ such that, for $N > N_c$*

$$|\mathbb{E}^{\pi^N} \hat{f}_{N,a,\delta} - \mathbb{E}^\pi f| \leq \epsilon.$$

Thus the computational cost grows to achieve this error grows with N as $N \times C(N)$.

PROOF. The proof is similar to that of Theorem 4.1, but using the limit Theorem 4.2 in place of Theorem 3.2. As a consequence we have $\delta = \hat{\ell}/N$ and, since $K = T/\delta$ we obtain $K = \lfloor TN/\hat{\ell} \rfloor$. \square

Thus we see that, even if one tunes the variance of RWM ‘optimally’ (in the sense of optimal scaling) the computational gain is negligible when compared to the gain obtained by tuning the mean to obtain the pCN algorithm. This reinforces the take home message of the paper, namely that for nonparametric inference, the design of tailored proposals, which take into account infinite dimensional structure, can have significant impact on computational complexity.

5. Key Estimates. This section assembles various results which are used in the previous section. Some of the technical proofs are deferred to the appendix.

5.1. Acceptance Probability Asymptotics. This section describes a first order expansion of the acceptance probability. The approximation

$$\alpha^\delta(x, \xi) \approx \bar{\alpha}^\delta(x, \xi) \quad \text{where} \quad \bar{\alpha}^\delta(x, \xi) = 1 - \sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle \mathbb{I}_{\{\langle \nabla \Psi(x), \xi \rangle > 0\}} \quad (5.1)$$

is valid for $\delta \ll 1$. The quantity $\bar{\alpha}^\delta$ has the advantage over α^δ of being very simple to analyse: explicit computations are available. This will be exploited in section 5.2. The quality of the approximation (5.1) is rigorously quantified in the next lemma.

LEMMA 5.1. (Acceptance probability estimate)

Let Assumptions 2.1 hold. For any integer $p \geq 1$ the quantity $\bar{\alpha}^\delta(x, \xi)$ satisfies

$$\mathbb{E}_x[|\alpha^\delta(x, \xi) - \bar{\alpha}^\delta(x, \xi)|^p] \lesssim \delta^p (1 + \|x\|_s^{2p}). \quad (5.2)$$

PROOF. See Appendix A. \square

Recall the local mean acceptance probability defined by $\alpha^\delta(x) = \mathbb{E}_x[\alpha^\delta(x, \xi)]$ in Equation (3.5). Define the approximate local mean acceptance probability by $\bar{\alpha}^\delta(x) \stackrel{\text{def}}{=} \mathbb{E}_x[\bar{\alpha}^\delta(x, \xi)]$. We now use Lemma 5.1 to approximate the local mean acceptance probability $\alpha^\delta(x)$.

COROLLARY 5.2. *Let Assumptions 2.1 hold. For any integer $p \geq 1$ the following estimates hold,*

$$|\alpha^\delta(x) - \bar{\alpha}^\delta(x)| \lesssim \delta (1 + \|x\|_s^2) \quad (5.3)$$

$$\mathbb{E}_x \left[|\alpha^\delta(x, \xi) - 1|^p \right] \lesssim \delta^{\frac{p}{2}} (1 + \|x\|_s^p) \quad (5.4)$$

PROOF. See Appendix A. \square

5.2. *Drift Estimates.* Then next lemma shows that explicit computations are available for the quantity $\bar{\alpha}^\delta$. We will use these explicit computations, together with quantification of the error committed in replacing α^δ by $\bar{\alpha}^\delta$, to estimate the mean drift (in this section) and the diffusion term (in the next section).

LEMMA 5.3. *The approximate acceptance probability $\bar{\alpha}^\delta(x, \xi)$ satisfies*

$$\sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x \left[\bar{\alpha}^\delta(x, \xi) \cdot \xi \right] = -C \nabla \Psi(x) \quad \forall x \in \mathcal{H}^s.$$

PROOF. Let $u = \sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x \left[\bar{\alpha}^\delta(x, \xi) \cdot \xi \right] \in \mathcal{H}^s$. To prove the lemma it suffices to verify that for all $v \in \mathcal{H}^{-s}$ we have

$$\langle u, v \rangle = -\langle C \nabla \Psi(x), v \rangle.$$

To this end, use the decomposition $v = \alpha \nabla \Psi(x) + w$ where $\alpha \in \mathbb{R}$ and $w \in \mathcal{H}^{-s}$ satisfies $\langle C \nabla \Psi(x), w \rangle = 0$. Since $\xi \stackrel{\mathcal{D}}{\sim} N(0, C)$ the two Gaussian random variables

$$Z_\Psi \stackrel{\text{def}}{=} \langle \nabla \Psi(x), \xi \rangle \quad \text{and} \quad Z_w \stackrel{\text{def}}{=} \langle w, \xi \rangle$$

are independent: indeed, (Z_Ψ, Z_w) is a Gaussian vector in \mathbb{R}^2 with $\text{Cov}(Z_\Psi, Z_w) = 0$. It thus follows that

$$\begin{aligned} \langle u, v \rangle &= -2 \langle \mathbb{E}_x [\langle \nabla \Psi(x), \xi \rangle 1_{\{\langle \nabla \Psi(x), \xi \rangle > 0\}} \cdot \xi] , \alpha \nabla \Psi(x) + w \rangle \\ &= -2 \mathbb{E}_x \left[\alpha Z_\Psi^2 1_{\{Z_\Psi > 0\}} + Z_w Z_\Psi 1_{\{Z_\Psi > 0\}} \right] \\ &= -2\alpha \mathbb{E}_x \left[Z_\Psi^2 1_{\{Z_\Psi > 0\}} \right] = -\alpha \mathbb{E}_x \left[Z_\Psi^2 \right] \\ &= -\alpha \langle C \nabla \Psi(x), \nabla \Psi(x) \rangle = \langle -C \nabla \Psi(x), \alpha \nabla \Psi(x) + w \rangle \\ &= -\langle C \nabla \Psi(x), v \rangle, \end{aligned}$$

which concludes the proof of Lemma 5.3. \square

We now use this explicit computation to give a proof of the drift estimate Lemma 3.6.

PROOF OF LEMMA 3.6. The function d^δ defined by Equation (3.8) can also be expressed as

$$d^\delta(x) = \left\{ \frac{(1-2\delta)^{\frac{1}{2}} - 1}{\delta} \alpha^\delta(x) x \right\} + \left\{ \sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x[\alpha^\delta(x, \xi) \xi] \right\} = B_1 + B_2, \quad (5.5)$$

where the mean local acceptance probability $\alpha^\delta(x)$ has been defined in Equation (3.5) and the two terms B_1 and B_2 are studied below. To prove Equation (3.20), it suffices to establish that

$$\|B_1 + x\|_s^p \lesssim \delta^{\frac{p}{2}} (1 + \|x\|_s^{2p}) \quad \text{and} \quad \|B_2 + C\nabla\Psi(x)\|_s^p \lesssim \delta^{\frac{p}{2}} (1 + \|x\|_s^{2p}). \quad (5.6)$$

We now establish these two bounds.

- Lemma 5.1 and Corollary 5.2 show that

$$\begin{aligned} \|B_1 + x\|_s^p &= \left\{ \frac{(1-2\delta)^{\frac{1}{2}} - 1}{\delta} \alpha^\delta(x) + 1 \right\}^p \|x\|_s^p \\ &\lesssim \left\{ \left| \frac{(1-2\delta)^{\frac{1}{2}} - 1}{\delta} - 1 \right|^p + |\alpha^\delta(x) - 1|^p \right\} \|x\|_s^p \\ &\lesssim \left\{ \delta^p + \delta^{\frac{p}{2}} (1 + \|x\|_s^p) \right\} \|x\|_s^p \lesssim \delta^{\frac{p}{2}} (1 + \|x\|_s^{2p}). \end{aligned} \quad (5.7)$$

- Lemma 5.1 shows that

$$\begin{aligned} \|B_2 + C\nabla\Psi(x)\|_s^p &= \left\| \sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x[\alpha^\delta(x, \xi) \xi] + C\nabla\Psi(x) \right\|_s^p \\ &\lesssim \delta^{-\frac{p}{2}} \left\| \mathbb{E}_x[\{\alpha^\delta(x, \xi) - \bar{\alpha}^\delta(x, \xi)\} \xi] \right\|_s^p + \underbrace{\left\| \sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x[\bar{\alpha}^\delta(x, \xi) \xi] + C\nabla\Psi(x) \right\|_s^p}_{=0}. \end{aligned} \quad (5.8)$$

By Lemma 5.3, the second term on the right hand is equal to zero. Consequently, Cauchy Schwarz' inequality implies that

$$\begin{aligned} \|B_2 + C\nabla\Psi(x)\|_s^p &\lesssim \delta^{-\frac{p}{2}} \mathbb{E}_x[|\alpha^\delta(x, \xi) - \bar{\alpha}^\delta(x, \xi)|^2]^{\frac{p}{2}} \\ &\lesssim \delta^{-\frac{p}{2}} \left(\delta^2 (1 + \|x\|_s^4) \right)^{\frac{p}{2}} \lesssim \delta^{\frac{p}{2}} (1 + \|x\|_s^{2p}). \end{aligned}$$

Estimates (5.7) and (5.8) give Equation (5.6). To complete the proof we establish the bound (3.21). The expression (5.5) shows that it suffices to verify

$$\sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x[\alpha^\delta(x, \xi) \xi] \lesssim 1 + \|x\|_s.$$

To this end, we use Lemma 5.3 and Corollary 5.2. By Cauchy-Schwarz,

$$\left\| \sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x[\alpha^\delta(x, \xi) \cdot \xi] \right\|_s = \left\| \sqrt{\frac{2\tau}{\delta}} \mathbb{E}_x[(\alpha^\delta(x, \xi) - 1) \cdot \xi] \right\|_s \lesssim \delta^{-\frac{1}{2}} \mathbb{E}_x[(\alpha^\delta(x, \xi) - 1)^2]^{\frac{1}{2}} \lesssim 1 + \|x\|_s,$$

which concludes the proof of Lemma 3.6. \square

5.3. *Noise Estimates.* In this section we estimate the error in the approximation $\Gamma^{k,\delta} \approx N(0, C_s)$. To this end, let us introduce the covariance operator $D^\delta(x)$ of the martingale difference Γ^δ ,

$$D^\delta(x) = \mathbb{E} \left[\Gamma^{k,\delta} \otimes_{\mathcal{H}^s} \Gamma^{k,\delta} \mid x^{k,\delta} = x \right].$$

For any $x, u, v \in \mathcal{H}^s$ the operator $D^\delta(x)$ satisfies

$$\mathbb{E} \left[\langle \Gamma^{k,\delta}, u \rangle_s \langle \Gamma^{k,\delta}, v \rangle_s \mid x^{k,\delta} = x \right] = \langle u, D^\delta(x) v \rangle_s.$$

The next lemma gives a quantitative version of the approximation $D^\delta(x) \approx C_s$.

LEMMA 5.4. (Noise estimates)

Let Assumptions 2.1 hold. For any pair of indices $i, j \geq 1$, the martingale difference term $\Gamma^\delta(x, \xi)$ satisfies

$$|\langle \hat{\varphi}_i, D^\delta(x) \hat{\varphi}_j \rangle_s - \langle \hat{\varphi}_i, C_s \hat{\varphi}_j \rangle_s| \lesssim \delta^{\frac{1}{8}} \cdot (1 + \|x\|_s) \quad (5.9)$$

$$|\text{Trace}_{\mathcal{H}^s}(D^\delta(x)) - \text{Trace}_{\mathcal{H}^s}(C_s)| \lesssim \delta^{\frac{1}{8}} \cdot (1 + \|x\|_s^2). \quad (5.10)$$

PROOF. See Appendix A. □

5.4. *A Priori Bound.* Now we have all the ingredients for the proof of the a priori bound presented in Lemma 3.8 which states that the rescaled process z^δ given by Equation (1.7) does not blow up in finite time.

PROOF LEMMA 3.8. Without loss of generality, assume that $p = 2n$ for some positive integer $n \geq 1$. We now prove that there exist constants $\alpha_1, \alpha_2, \alpha_3 > 0$ satisfying

$$\mathbb{E}[\|x^{k,\delta}\|_s^{2n}] \leq (\alpha_1 + \alpha_2 k \delta) e^{\alpha_3 k \delta}. \quad (5.11)$$

Lemma 3.8 is a straightforward consequence of Equation 5.11 since this implies that

$$\delta \sum_{k\delta < T} \mathbb{E}[\|x^{k,\delta}\|_s^{2n}] \leq \delta \sum_{k\delta < T} (\alpha_1 + \alpha_2 k \delta) e^{\alpha_3 k \delta} \asymp \int_0^T (\alpha_1 + \alpha_2 t) e^{\alpha_3 t} dt < \infty.$$

For notational convenience, let us define $V^{k,\delta} = \mathbb{E}[\|x^{k,\delta}\|_s^{2n}]$. To prove Equation (5.11), it suffices to establish that

$$V^{k+1,\delta} - V^{k,\delta} \leq K \delta \cdot (1 + V^{k,\delta}), \quad (5.12)$$

where $K > 0$ is constant independent from $\delta \in (0, \frac{1}{2})$. Indeed, iterating inequality (5.12) leads to the bound (5.11), for some computable constants $\alpha_1, \alpha_2, \alpha_3 > 0$. The definition of V^k shows that

$$\begin{aligned} V^{k+1,\delta} - V^{k,\delta} &= \mathbb{E}[\|x^{k,\delta} + (x^{k+1,\delta} - x^{k,\delta})\|_s^{2n} - \|x^{k,\delta}\|_s^{2n}] \\ &= \mathbb{E} \left[\left\{ \|x^{k,\delta}\|_s^2 + \|x^{k+1,\delta} - x^{k,\delta}\|_s^2 + 2 \langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta} \rangle_s \right\}^n - \|x^{k,\delta}\|_s^{2n} \right] \end{aligned} \quad (5.13)$$

where the increment $x^{k+1,\delta} - x^{k,\delta}$ is given by

$$x^{k+1,\delta} - x^{k,\delta} = \gamma^{k,\delta} \left((1 - 2\delta)^{\frac{1}{2}} - 1 \right) x^{k,\delta} + \sqrt{2\delta} \gamma^{k,\delta} \xi^k. \quad (5.14)$$

To bound the right-hand-side of Equation (5.13), we use a binomial expansion and control each term. To this end, we establish the following estimate: for all integers $i, j, k \geq 0$ satisfying

$$i + j + k = n \quad \text{and} \quad (i, j, k) \neq (n, 0, 0)$$

the following inequality holds,

$$\mathbb{E} \left[\left(\|x^{k,\delta}\|_s^2 \right)^i \left(\|x^{k+1,\delta} - x^{k,\delta}\|_s^2 \right)^j \left(\langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta} \rangle_s \right)^k \right] \lesssim \delta (1 + V^{k,\delta}). \quad (5.15)$$

To prove Equation (5.15), we separate two different cases.

- Let us suppose $(i, j, k) = (n-1, 0, 1)$. Lemma 3.6 states that the approximate drift has a linearly bounded growth so that

$$\left\| \mathbb{E} [x^{k+1,\delta} - x^{k,\delta} | x^{k,\delta}] \right\|_s = \delta \|d^\delta(x^{k,\delta})\|_s \lesssim \delta (1 + \|x^{k,\delta}\|_s).$$

Consequently, we have

$$\begin{aligned} \mathbb{E} \left[\left(\|x^{k,\delta}\|_s^2 \right)^{n-1} \langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta} \rangle_s \right] &\lesssim \mathbb{E} \left[\|x^{k,\delta}\|_s^{2(n-1)} \|x^{k,\delta}\|_s \left(\delta (1 + \|x^{k,\delta}\|_s) \right) \right] \\ &\lesssim \delta (1 + V^{k,\delta}). \end{aligned}$$

This proves Equation (5.15) in the case $(i, j, k) = (n-1, 0, 1)$.

- Let us suppose $(i, j, k) \notin \{(n, 0, 0), (n-1, 0, 1)\}$. Because for any integer $p \geq 1$,

$$\mathbb{E}_x \left[\|x^{k+1,\delta} - x^{k,\delta}\|_s^p \right]^{\frac{1}{p}} \lesssim \delta^{\frac{1}{2}} (1 + \|x\|_s)$$

it follows from Cauchy-Schwarz inequality that

$$\mathbb{E} \left[\left(\|x^{k,\delta}\|_s^2 \right)^i \left(\|x^{k+1,\delta} - x^{k,\delta}\|_s^2 \right)^j \left(\langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta} \rangle_s \right)^k \right] \lesssim \delta^{j+\frac{k}{2}} (1 + V^{k,\delta}).$$

Since we have supposed that $(i, j, k) \notin \{(n, 0, 0), (n-1, 0, 1)\}$ and $i + j + k = n$, it follows that $j + \frac{k}{2} \geq 1$. This concludes the proof of Equation (5.15),

The binomial expansion of Equation (5.13) and the bound (5.15) show that

$$V^{k+1,\delta} - V^{k,\delta} \lesssim \delta (1 + V^{k,\delta}).$$

This proves Equation (5.12), which concludes the proof of Lemma 3.8. \square

5.5. Invariance Principle. Combining the noise estimates of Lemma 5.4 and the a priori bound of Lemma 3.8, we show that under Assumptions 2.1 the sequence of rescaled noise processes defined in Equation 3.9 converges weakly to a Brownian motion. This is the content of Lemma 3.7 whose proof is now presented.

PROOF OF LEMMA 3.7. As described in [Ber86] [Proposition 5.1], in order to prove that W^δ converges weakly to W in $C([0, T]; \mathcal{H}^s)$ it suffices to prove that for any $t \in [0, T]$ and any pair of indices $i, j \geq 0$ the following three limits hold in probability,

$$(5.16) \quad \lim_{\delta \rightarrow 0} \delta \sum_{k\delta < t} \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^2 |x^{k,\delta} \right] = t \cdot \text{Trace}_{\mathcal{H}^s}(C_s)$$

$$(5.17) \quad \lim_{\delta \rightarrow 0} \delta \sum_{k\delta < t} \mathbb{E} \left[\langle \Gamma^{k,\delta}, \hat{\varphi}_i \rangle_s \langle \Gamma^{k,\delta}, \hat{\varphi}_j \rangle_s |x^{k,\delta} \right] = t \langle \hat{\varphi}_i, C_s \hat{\varphi}_j \rangle_s$$

$$(5.18) \quad \lim_{\delta \rightarrow 0} \delta \sum_{k\delta < T} \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^2 \mathbb{I}_{\{\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1} \epsilon\}} |x^{k,\delta} \right] = 0 \quad \forall \epsilon > 0.$$

We now check that these three conditions are indeed satisfied.

- Condition (5.16): since $\mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^2 |x^{k,\delta} \right] = \text{Trace}_{\mathcal{H}^s}(D^\delta(x^{k,\delta}))$, Lemma 5.4 shows that

$$\mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^2 |x^{k,\delta} \right] = \text{Trace}_{\mathcal{H}^s}(C_s) + \mathbf{e}_1^\delta(x^{k,\delta})$$

where the error term \mathbf{e}_1^δ satisfies $|\mathbf{e}_1^\delta(x)| \lesssim \delta^{\frac{1}{8}} (1 + \|x\|_s^2)$. Consequently, to prove condition (5.16) it suffices to establish that

$$\lim_{\delta \rightarrow 0} \mathbb{E} \left[\delta \sum_{k\delta < T} \mathbf{e}_1^\delta(x^{k,\delta}) \right] = 0.$$

We have $\mathbb{E} \left[\delta \sum_{k\delta < T} \mathbf{e}_1^\delta(x^{k,\delta}) \right] \lesssim \delta^{\frac{1}{8}} \left\{ \delta \cdot \mathbb{E} \left[\sum_{k\delta < T} (1 + \|x^{k,\delta}\|_s^2) \right] \right\}$ and the apriori bound presented in Lemma 3.8 shows that

$$\sup_{\delta \in (0, \frac{1}{2})} \left\{ \delta \cdot \mathbb{E} \left[\sum_{k\delta < T} (1 + \|x^{k,\delta}\|_s^2) \right] \right\} < \infty.$$

Consequently $\lim_{\delta \rightarrow 0} \mathbb{E} \left[\delta \sum_{k\delta < T} \mathbf{e}_1^\delta(x^{k,\delta}) \right] = 0$, and the conclusion follows.

- Condition (5.17): Lemma 5.4 states that

$$\mathbb{E}_k \left[\langle \Gamma^{k,\delta}, \hat{\varphi}_i \rangle_s \langle \Gamma^{k,\delta}, \hat{\varphi}_j \rangle_s \right] = \langle \hat{\varphi}_i, C_s \hat{\varphi}_j \rangle_s + \mathbf{e}_2^\delta(x^{k,\delta})$$

where the error term \mathbf{e}_2^δ satisfies $|\mathbf{e}_2^\delta(x)| \lesssim \delta^{\frac{1}{8}} (1 + \|x\|_s)$. The exact same approach as the proof of Condition (5.16) gives the conclusion.

- Condition (5.18): from Cauchy-Schwarz and Markov's inequalities it follows that

$$\begin{aligned} \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^2 \mathbb{I}_{\{\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1} \epsilon\}} \right] &\leq \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^4 \right]^{\frac{1}{2}} \cdot \mathbb{P} \left[\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1} \epsilon \right]^{\frac{1}{2}} \\ &\leq \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^4 \right]^{\frac{1}{2}} \cdot \left\{ \frac{\mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^4 \right]}{(\delta^{-1} \epsilon)^2} \right\}^{\frac{1}{2}} \\ &\leq \frac{1}{\epsilon^2} \delta^2 \cdot \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^4 \right]. \end{aligned}$$

Consequently we have

$$\mathbb{E} \left[\delta \sum_{k\delta < T} \mathbb{E} \left[\|\Gamma^{k,\delta}\|_s^2 \mathbb{I}_{\{\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1} \epsilon\}} |x^{k,\delta} \right] \right] \leq \frac{1}{\epsilon^2} \delta^2 \left\{ \delta \cdot \mathbb{E} \left[\sum_{k\delta < T} \|\Gamma^{k,\delta}\|_s^4 \right] \right\}$$

and the conclusion again follows from the a priori bound Lemma 3.8.

□

6. Quadratic Variation. As discussed in the introduction, the SPDE (1.6), and the Metropolis-Hastings algorithm pCN which approximates it for small δ , do not satisfy the smoothing property and so almost sure properties of the limit measure π^τ are not necessarily seen at finite time. In this section we prove a limit theorem satisfied by such almost sure quantities, under pCN. When C is the covariance of Brownian motion or Brownian bridge then the almost sure quantity will be quadratic variation; for other covariances it will be a generalization defined precisely in the following subsection. We show that the quadratic variation of pCN converges as $k \rightarrow \infty$ to its value under the invariant measure. We then prove that piecewise linear interpolation of this quantity solves, in the small δ limit, a linear ODE (the “fluid limit”) whose globally attractive stable state is the almost sure quantity. This quantifies the manner in which the pCN method approaches statistical equilibrium.

6.1. Definition and Properties. Under Assumptions 2.1, the Karhunen-Loève expansion shows that π_0 -almost every $x \in \mathcal{H}$ satisfies

$$\lim_{N \rightarrow \infty} N^{-1} \sum_{j=1}^N \frac{\langle x, \varphi_j \rangle^2}{\lambda_j^2} = 1.$$

This motivates the definition of the quadratic variation like quantities

$$V_-(x) \stackrel{\text{def}}{=} \liminf_{N \rightarrow \infty} N^{-1} \sum_{j=1}^n \frac{\langle x, \varphi_j \rangle^2}{\lambda_j^2} \quad \text{and} \quad V_+(x) \stackrel{\text{def}}{=} \limsup_{N \rightarrow \infty} N^{-1} \sum_{j=1}^n \frac{\langle x, \varphi_j \rangle^2}{\lambda_j^2}.$$

When these two quantities are equal the vector $x \in \mathcal{H}$ is said to possess a *quadratic variation* $V(x)$ defined as $V(x) = V_-(x) = V_+(x)$. Consequently, π_0 -almost every $x \in \mathcal{H}$ possesses a quadratic variation $V(x) = 1$. It is a straightforward consequence that π_0^τ -almost every and π^τ -almost every $x \in \mathcal{H}$ possesses a quadratic variation $V(x) = \tau$. Strictly speaking this only coincides with quadratic variation when C is the covariance of a (possibly conditioned) Brownian motion; however we use the terminology more generally in this section. The next lemma proves that the quadratic variation $V(\cdot)$ behaves as it should do with respect to additivity.

LEMMA 6.1. (Quadratic Variation Additivity)

Consider a vector $x \in \mathcal{H}$ and a Gaussian random variable $\xi \stackrel{\mathcal{D}}{\sim} \pi_0$ and a real number $\alpha \in \mathbb{R}$. Suppose that the vector $x \in \mathcal{H}$ possesses a finite quadratic variation $V(x) < +\infty$. Then almost surely the vector $x + \alpha\xi \in \mathcal{H}$ possesses a quadratic variation that is equal to

$$V(x + \alpha\xi) = V(x) + \alpha^2.$$

PROOF. Let us define $V_N \stackrel{\text{def}}{=} N^{-1} \sum_1^N \frac{\langle x, \varphi_j \rangle \cdot \langle \xi, \varphi_j \rangle}{\lambda_j^2}$. To prove Lemma 6.1 it suffices to prove that almost surely the following limit holds

$$\lim_{N \rightarrow \infty} V_N = 0.$$

Borel-Cantelli Lemma shows that it suffices to prove that for every fixed $\varepsilon > 0$ we have $\sum_{N \geq 1} \mathbb{P}[|V_N| > \varepsilon] < \infty$. Notice then that V_N is a centred Gaussian random variables with variance

$$\text{Var}(V_N) = \frac{1}{N} \left(N^{-1} \sum_1^N \frac{\langle x, \varphi_j \rangle^2}{\lambda_j^2} \right) \asymp \frac{V(x)}{N}.$$

It readily follows that $\sum_{N \geq 1} \mathbb{P}[|V_N| > \varepsilon] < \infty$, finishing the proof of the Lemma. \square

6.2. *Large k Behaviour of Quadratic Variation for pCN.* The pCN algorithm at temperature $\tau > 0$ and discretization parameter $\delta > 0$ proposes a move from x to y according to the dynamics

$$y = (1 - 2\delta)^{\frac{1}{2}} x + (2\delta\tau)^{\frac{1}{2}} \xi \quad \text{with} \quad \xi \stackrel{\mathcal{D}}{\sim} \pi_0.$$

This move is accepted with probability $\alpha^\delta(x, y)$. In this case, Lemma 6.1 shows that if the quadratic variation $V(x)$ exists then the quadratic variation of the proposed move $y \in \mathcal{H}$ exists and satisfies

$$\frac{V(y) - V(x)}{\delta} = -2(V(x) - \tau). \quad (6.1)$$

Consequently, one can prove that for any finite time step $\delta > 0$ and temperature $\tau > 0$ the quadratic variation of the MCMC algorithm converges to τ .

PROPOSITION 6.2. (Limiting Quadratic Variation) *Let Assumptions 2.1 hold and $\{x^{k,\delta}\}_{k \geq 0}$ be the Markov chain of section 3.1. Then almost surely the quadratic variation of the Markov chain converges to τ ,*

$$\lim_{k \rightarrow \infty} V(x^{k,\delta}) = \tau.$$

PROOF. Let us first show that the number of accepted moves is infinite. If this were not the case, the Markov chain would eventually reach a position $x^{k,\delta} = x \in \mathcal{H}$ such that all subsequent proposals $y^{k+l} = (1 - 2\delta)^{\frac{1}{2}} x^k + (2\delta\tau)^{\frac{1}{2}} \xi^{k+l}$ would be refused. This means that the i.i.d. Bernoulli random variables $\gamma^{k+l} = \text{Bernoulli}(\alpha^\delta(x^k, y^{k+l}))$ satisfy $\gamma^{k+l} = 0$ for all $l \geq 0$. This can only happen with probability 0. Indeed, since $\mathbb{P}[\gamma^{k+l} = 1] > 0$, one can use Borel-Cantelli Lemma to show that almost surely there exists $l \geq 0$ such that $\gamma^{k+l} = 1$. To conclude the proof of the Proposition, notice then that the sequence $\{u_k\}_{k \geq 0}$ defined by $u_{k+1} - u_k = -2\delta(u_k - \tau)$ converges to τ . \square

6.3. *Fluid Limit for Quadratic Variation of pCN.* To gain further insight into the rate at which the limiting behaviour of the quadratic variation is observed for pCN we derive an ODE “fluid limit” for the Metropolis-Hastings algorithm. We introduce the continuous time process $t \mapsto v^\delta(t)$ defined as continuous piecewise linear interpolation of the the process $k \mapsto V(x^{k,\delta})$ as follows:

$$v^\delta(t) = \frac{1}{\delta} (t - t_k) V(x^{k+1,\delta}) + \frac{1}{\delta} (t_{k+1} - t) V(x^{k,\delta}) \quad \text{for} \quad t_k \leq t < t_{k+1}. \quad (6.2)$$

Since the acceptance probability of pCN approaches 1 as $\delta \rightarrow 0$ (see Corollary 5.2) Equation (6.1) shows heuristically that the trajectories of the process $t \mapsto v^\delta(t)$ should be well approximated by the solution of the (non stochastic) differential equation

$$\dot{v} = -2(v - \tau) \quad (6.3)$$

We prove such a result, in the sense of convergence in probability in $C([0, T]; \mathbb{R})$:

THEOREM 6.3. (Theorem 3: Fluid Limit For Quadratic Variation) *Let Assumptions 2.1 hold. Let the Markov chain x^δ start at fixed position $x_* \in \mathcal{H}^s$. Assume that $x_* \in \mathcal{H}$ possesses a finite quadratic variation, $V(x_*) < \infty$. Then the function $v^\delta(t)$ converges in probability in $C([0, T], \mathbb{R})$, as δ goes to 0, to the solution of the differential equation (6.3) with initial condition $v_0 = V(x_*)$.*

As already indicated, the heart of the proof of the result consists in showing that the acceptance probability of the algorithm converges to 1 as δ goes to 0. We prove such a result as Lemma 6.4 below, and then proceed to prove Theorem 6.3. To this end we introduce $t^\delta(k)$, the number of accepted moves:

$$t^\delta(k) \stackrel{\text{def}}{=} \sum_{l \leq k} \gamma^{l,\delta},$$

where $\gamma^{l,\delta} = \text{Bernoulli}(\alpha^\delta(x, y))$ is the Bernoulli random variable defined in Equation (3.4). Since the acceptance probability of the algorithm converges to 1 as $\delta \rightarrow 0$, the approximation $t^\delta(k) \approx k$ holds. In order to prove a fluid limit result on the interval $[0, T]$ one needs to prove that the quantity $|t^\delta(k) - k|$ is small when compared to δ^{-1} . The next Lemma shows that such a bounds holds uniformly on the interval $[0, T]$.

LEMMA 6.4. (Number of Accepted Moves) *Let Assumptions 2.1 hold. The number of accepted moves $t^\delta(\cdot)$ verifies*

$$\lim_{\delta \rightarrow 0} \sup \{ \delta \cdot |t^\delta(k) - k| : 0 \leq k \leq T\delta^{-1} \} = 0$$

where the convergence holds in probability.

PROOF. The proof is given in Appendix B. □

We now complete the proof of Theorem 6.3 using the key Lemma 6.4.

PROOF OF THEOREM 6.3. The proof consists in proving that the trajectory of the quadratic variation process behaves as if all the move were accepted. The main ingredient is the uniform lower bound on the acceptance probability given by Lemma 6.4.

Recall that $v^\delta(k\delta) = V(x^{k,\delta})$. Consider the piecewise linear function $\hat{v}^\delta(\cdot) \in C([0, T], \mathbb{R})$ defined by linear interpolation of the values $\hat{v}^\delta(k\delta) = u^\delta(k)$ and where the sequence $\{u^\delta(k)\}_{k \geq 0}$ satisfies $u^\delta(0) = V(x_*)$ and

$$u^\delta(k+1) - u^\delta(k) = -2\delta(u^\delta(k) - \tau).$$

The value $u^\delta(k) \in \mathbb{R}$ represents the quadratic variation of $x^{k,\delta}$ if the k first moves of the MCMC algorithm had been accepted. One can readily check that as δ goes to zero the sequence of continuous functions $\hat{v}^\delta(\cdot)$ converges in $C([0, T], \mathbb{R})$ to the solution $v(\cdot)$ of the differential equation (6.3). Consequently, to prove Theorem 6.3 it suffices to show that for any $\varepsilon > 0$ we have

$$\lim_{\delta \rightarrow 0} \mathbb{P} \left[\sup \left\{ |V(x^{k,\delta}) - u^\delta(k)| : k \leq \delta^{-1}T \right\} > \varepsilon \right] = 0. \quad (6.4)$$

The definition of the number of accepted moves $t^\delta(k)$ is such that $V(x^{k,\delta}) = u^\delta(t^\delta(k))$. Note that

$$(6.5) \quad u^\delta(k) = (1 - 2\delta)^k u_0 + (1 - (1 - 2\delta)^k) \tau.$$

Hence, for any integers $t_1, t_2 \geq 0$, we have $|u^\delta(t_2) - u^\delta(t_1)| \leq |u^\delta(|t_2 - t_1|) - u^\delta(0)|$ so that

$$|V(x^{k,\delta}) - u^\delta(k)| = |u^\delta(t^\delta(k)) - u^\delta(k)| \leq |u^\delta(k - t^\delta(k)) - u^\delta(0)|.$$

Equation (6.5) shows that $|u^\delta(k) - u^\delta(0)| \lesssim (1 - (1 - 2\delta)^k)$. This implies that

$$|V(x^{k,\delta}) - u^\delta(k)| \lesssim 1 - (1 - 2\delta)^{k-t^\delta(k)} \lesssim 1 - (1 - 2\delta)^{\delta^{-1}S}$$

where $S = \sup \{\delta \cdot |t^\delta(k) - k| : 0 \leq k \leq T\delta^{-1}\}$. Since for any $a > 0$ we have $1 - (1 - 2\delta)^{a\delta^{-1}} \rightarrow 1 - e^{-2a}$, Equation (6.4) follows if one can prove that as δ goes to 0 the supremum S converges to 0 in probability: this is precisely the content of Lemma 6.4. This concludes the proof of Theorem 6.3. \square

7. Numerical Results. The numerical results comparing the efficiency of the RWM and pCN are presented in a companion paper [CRSW11] and we refer the interested reader to that paper. In this section, we present some numerical simulations demonstrating our results in the context of simulated annealing. We consider the minimisation of a functional $J(\cdot)$ defined on the Sobolev space $H_0^1(\mathbb{R}) \subset C^0([0, 1]) \subset L^2(0, 1)$. Functions $x \in H_0^1([0, 1])$ are continuous and satisfy $x(0) = x(1) = 0$. For a given real parameter $\lambda > 0$, the functional $J : H_0^1([0, 1]) \rightarrow \mathbb{R}$ is composed of two competitive terms, as follows:

$$J(x) = \frac{1}{2} \int_0^1 |\dot{x}(s)|^2 ds + \frac{\lambda}{4} \int_0^1 (x(s)^2 - 1)^2 ds. \quad (7.1)$$

The first term penalises functions that deviate from being flat, whilst the second term penalises functions that deviate from one in absolute value. Critical points of the functional $J(\cdot)$ solve the following Euler-Lagrange equation:

$$\begin{aligned} \ddot{x} + \lambda x(1 - x^2) &= 0 \\ x(0) &= x(1) = 0. \end{aligned} \quad (7.2)$$

Clearly $x \equiv 0$ is a solution for all $\lambda \in \mathbb{R}^+$. If $\lambda \in (0, \pi^2)$ then this is the unique solution of the Euler-Lagrange equation and is the global minimizer of J . For each integer k there is a supercritical bifurcation at parameter value $\lambda = k^2\pi^2$. For $\lambda > \pi^2$ there are two minimizers, both of one sign and one being minus the other. The three different solutions of (7.2) which exist for $\lambda = 2\pi^2$ are displayed in Figure 1, at which value the zero (blue dotted) solution is a saddle point, and the two green solutions are the global minimizers of J . These properties of J are overviewed in, for example, [Hen81]. We will show how these global minimizers can emerge from an algorithm whose only ingredients are an ability to evaluate Ψ and to sample from the Gaussian measure with Cameron-Martin norm $\int_0^1 |\dot{x}(s)|^2 ds$. We emphasize that we are not advocating this as the optimal method for solving the Euler-Lagrange equations (7.2). We have chosen this example for its simplicity, in order to illustrate the key ingredients of the theory developed in this paper.

The pCN algorithm to minimize J given by (7.1) is implemented on $L^2([0, 1])$. Recall from section 1 that the Gaussian measure $N(0, C)$ may be identified by finding the covariance operator for which the $H_0^1([0, 1])$ norm $\|x\|_C^2 \stackrel{\text{def}}{=} \int_0^1 |\dot{x}(s)|^2 ds$ is the Cameron-Martin norm. In [HAVW05] it is shown that the Wiener bridge measure $\mathbb{W}_{0 \rightarrow 0}$ on $L^2([0, 1])$ has precisely this Cameron-Martin norm; indeed it is demonstrated that C^{-1} is the densely defined operator $-\frac{d^2}{ds^2}$ with $D(C^{-1}) = H^2([0, 1]) \cap H_0^1([0, 1])$. In this regard it is also instructive to adopt the physicists viewpoint that

$$\mathbb{W}_{0 \rightarrow 0}(dx) \propto \exp\left(-\frac{1}{2} \int_0^1 |\dot{x}(s)|^2 ds\right) dx$$

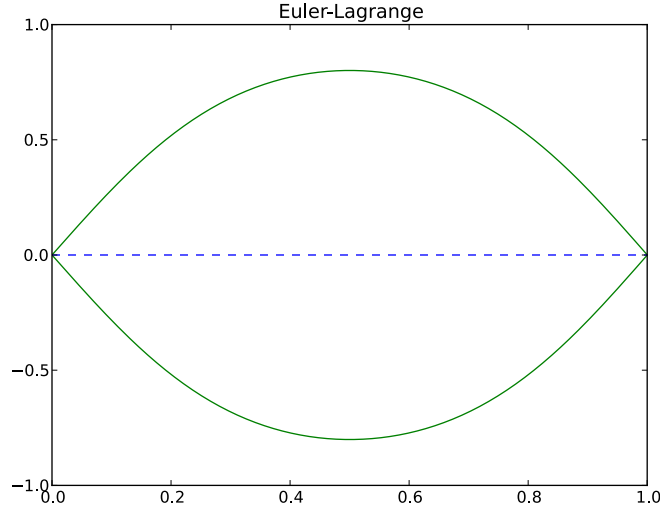


FIG 1. The three solutions of the Euler-Lagrange Equation (7.2) for $\lambda = 2\pi^2$. Only the two non-zero solutions are global minimum of the functional $J(\cdot)$. The dotted solution is a local maximum of $J(\cdot)$.

although, of course, there is no Lebesgue measure in infinite dimensions. Using an integration by parts, together with the boundary conditions on $H_0^1([0, 1])$, then gives

$$\mathbb{W}_{0 \rightarrow 0}(dx) \propto \exp\left(\frac{1}{2} \int_0^1 x(s) \frac{d^2 x}{ds^2}(s) ds\right) dx$$

and the inverse of C is clearly identified as the differential operator above. See [CH06] for basic discussion of the physicists viewpoint on Wiener measure. For a given temperature parameter τ the Wiener bridge measure $\mathbb{W}_{0 \rightarrow 0}^\tau$ on $L^2([0, 1])$ is defined as the law of $\{\sqrt{\tau} W(t)\}_{t \in [0, 1]}$ where $\{W(t)\}_{t \in [0, 1]}$ is a standard Brownian bridge on $[0, 1]$ drawn from $\mathbb{W}_{0 \rightarrow 0}$.

The posterior distribution $\pi^\tau(dx)$ is defined by the change of probability formula

$$\frac{d\pi^\tau}{d\mathbb{W}_{0 \rightarrow 0}^\tau}(x) \propto e^{-\Psi(x)} \quad \text{with} \quad \Psi(x) = \frac{\lambda}{4} \int_0^1 (x(s)^2 - 1)^2 ds.$$

Notice that $\pi_0^\tau(H_0^1([0, 1])) = \pi^\tau(H_0^1([0, 1])) = 0$ since a Brownian bridge is almost surely not differentiable anywhere on $[0, 1]$. It is for this reason that the algorithm is implemented on $L^2([0, 1])$ even though the functional $J(\cdot)$ is defined on the Sobolev space $H_0^1([0, 1])$. In terms of Assumptions 2.1(1) we have $\kappa = 1$ and the measure π_0^τ is supported on \mathcal{H}^r if and only if $r < \frac{1}{2}$, see Remark 2.2; note also that $H_0^1([0, 1]) = \mathcal{H}^1$. Assumption 2.1(2) is satisfied for any choice $s \in [\frac{1}{4}, \frac{1}{2})$ because \mathcal{H}^s is embedded into $L^4([0, 1])$ for $s \geq \frac{1}{4}$. We add here that Assumptions 2.1(3-4) do not hold globally, but only locally on bounded sets, but the numerical results below will indicate that the theory developed in this paper is still relevant and could be extended to nonlocal versions of Assumptions 2.1(3-4), with considerable further work.

Following section 3.1, the pCN Markov chain at temperature $\tau > 0$ and time discretization $\delta > 0$ proposes moves from x to y according to

$$y = (1 - 2\delta)^{\frac{1}{2}} x + (2\delta\tau)^{\frac{1}{2}} \xi$$

where $\xi \in C([0, 1], \mathbb{R})$ is a standard Brownian bridge on $[0, 1]$. The move $x \rightarrow y$ is accepted with probability $\alpha^\delta(x, \xi) = 1 \wedge \exp(-\tau^{-1}[\Psi(y) - \Psi(x)])$. Figure 2 displays the convergence of the Markov chain $\{x^{k, \delta}\}_{k \geq 0}$ to a minimiser of the functional $J(\cdot)$. Note that this convergence is not shown with respect to the space $H_0^1([0, 1])$ on which J is defined, but rather in $L^2([0, 1])$; indeed $J(\cdot)$ is almost surely infinite when evaluated at samples of the pCN algorithm, precisely because $\pi_0^\tau(H_0^1([0, 1])) = 0$, as discussed above.

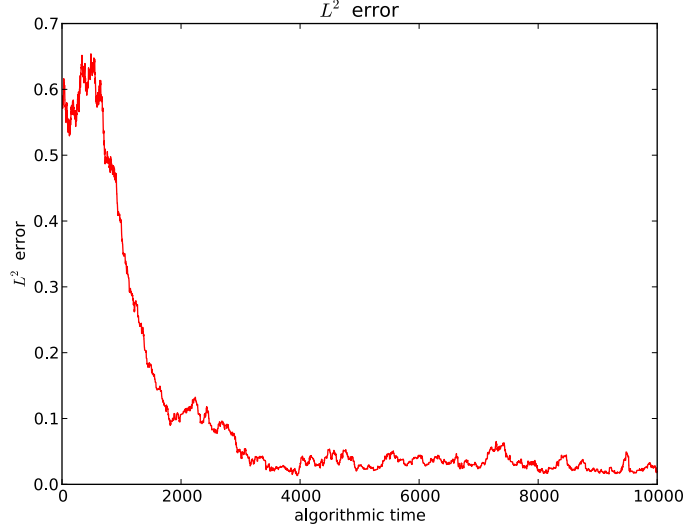


FIG 2. *pCN* parameters: $\lambda = 2\pi^2$, $\delta = 1.10^{-2}$, $\tau = 1.10^{-2}$. The algorithm is started at the zero function, $x^{0, \delta}(t) = 0$ for $t \in [0, 1]$. After a transient phase, the algorithm fluctuates around a global minimiser of functional $J(\cdot)$. The L^2 error $\|x^{k, \delta} - (\text{minimiser})\|_{L^2}$ is plotted as a function of the algorithmic time k .

Of course the algorithm does not converge *exactly* to a minimiser of $J(\cdot)$, but fluctuates in a neighbourhood of it. As described in the introduction of this article, in a finite dimensional setting the target probability distribution π^τ has Lebesgue density proportional to $\exp(-\tau^{-1}J(x))$. This intuitively shows that the size of the fluctuations around the minimum of the functional $J(\cdot)$ are of size proportional to $\sqrt{\tau}$. Figure 3 shows this phenomenon on log-log scales: the asymptotic mean error $\mathbb{E}[\|x - (\text{minimiser})\|_2]$ is displayed as a function of the temperature τ . Figure 4 illustrates Theorem 6.3. One can observe the path $\{v^\delta(t)\}_{t \in [0, T]}$ for a finite time step discretization parameter δ as well as the limiting path $\{v(t)\}_{t \in [0, T]}$ that is solution of the differential equation (6.3).

8. Conclusion. In the following we briefly summarize our results, their implications, the theoretical tools employed and scope for future work.

- The paper is organized around the **Optimal Proposal Design Principle** that proposals which are well-defined on the infinite dimensional parameter space results in MCMC methods which do not suffer from the curse of dimensionality. Our emphasis is on the general setting of Gaussian random field prior distributions and associated problems in Bayesian nonparametrics.
- We exhibit an example which highlights the importance of the optimal design of proposals in the context of random walk methods. We show that modifying the *mean* of the standard random walk proposal (RWM), to obtain the preconditioned Crank-Nicolson walk proposal

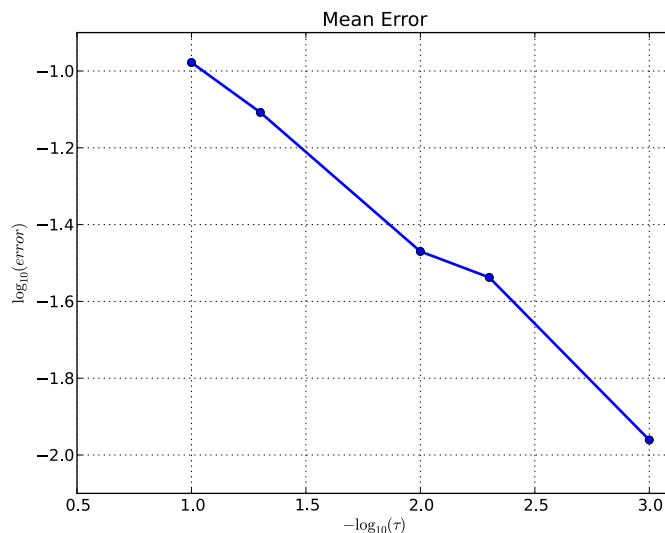


FIG 3.

Mean error $\mathbb{E}[\|x - (\text{minimiser})\|_2]$ as a function of the temperature τ .

(pCN), results in an order of magnitude improvement in computational complexity when applied to problems with Gaussian random field prior. In contrast, previous theories have concentrated on optimal scaling of the proposal *variance* [RGG97] and this has, in comparison, a negligible effect on complexity. Theorems 1 and 2 in the introduction show that when discretizing nonparametric problems to N dimensions the pCN method achieves a similar error to RWM with computational complexity reduced by a factor of order $\mathcal{O}(N)$.

- More generally, we emphasize the fact that designing MCMC algorithms for nonparametric problems needs more study since it is extremely useful for applied researchers to have guidelines regarding implementation of algorithms. For nonparametric problems, construction of prior distributions and likelihood modeling must be integrated with designing proposals and vice versa. This synthesis will have a beneficial impact on both the statistical inference and the efficiency of the algorithm. In this regard, we also advocate that in Bayesian nonparametric problems, the primary focus should be on optimal design of algorithms (which inherit the structure of prior and the likelihood) as compared to optimal scaling. A review of a wide range of algorithms which do this for Gaussian random field priors may be found in [CRSW11].
- Our basic tool of analysis is the construction of diffusion limits for the MCMC method and use of ergodicity of the limit process to quantify the computational cost of obtaining sample path averages which are close to the desired expectation. To prove these diffusion limits we have further developed the methods from [MPS11], which apply to the RWM, so that they may be used to study the pCN method. The diffusion limit results obtained in this paper use a proof technique analogous to the proof of diffusion limits of Markov chains in finite dimensions. We believe therefore that the methods of analysis that we introduce may be used to understand other MCMC algorithms, based on local proposals, and other nonparametric problems. The approach is encapsulated in Lemma 3.5 which is structured in such a way that it, or variants of it, may be used to prove diffusion limits for a variety of problems, especially when there is an underlying Gaussian structure.

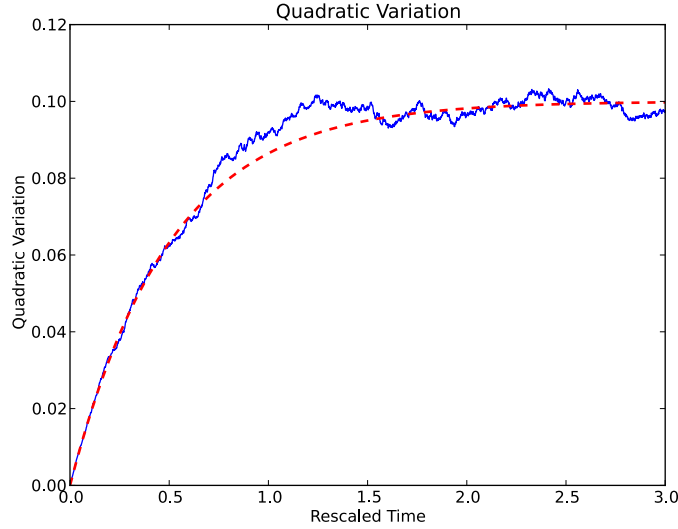


FIG 4. *pCN* parameters: $\lambda = 2\pi^2$, $\tau = 1.10^{-1}$, $\delta = 1.10^{-3}$ and the algorithm starts at $x^{k,\delta} = 0$. The rescaled quadratic variation process (full line) behaves as the solution of the differential equation (dotted line), as predicted by Theorem 6.3. The quadratic variation converges to τ , as described by Proposition 6.2.

- The recently developed analysis of 1-Wasserstein spectral gaps for MCMC in high dimensions [HSV11] shows great promise as a potential tool for the study of a wide range of MCMC methods developed to adhere to the design principle studied in this paper, even outside the class of local proposal MCMC methods. This is because the methods are well-adapted to the high dimensional limit, being based on study of infinite dimensional problems. In contrast, the methods of [MT93], which have been widely adopted by statisticians, do not scale so well with respect to dimension since they fail completely in many infinite dimensional settings.
- There are a host of tools known outside the statistics literature which are extremely useful for obtaining guidelines for implementation of algorithms. For instance, the *pCN* adheres to the “optimize then discretize” perspective whereas the RWM sticks to the “discretize then optimize” view point. In numerical analysis, it is known that in some problems, the former performs better than the latter (see [HPUU08], Chapter 3). We illustrate this point in statistical examples.
- We have demonstrated a class of algorithms to minimize the functional J given by (1.12). The Assumptions 2.1 encode the intuition that the quadratic part of J dominates. Under these assumptions we study the properties of an algorithm which requires only the evaluation of Ψ and the ability to draw samples from Gaussian measures with (Cameron-Martin) norm given by the quadratic part of J . We demonstrate that, in a certain parameter limit, the algorithm behaves like a noisy gradient flow for the functional J and that, furthermore, the size of the noise can be controlled systematically. Thus we have constructed a simulated annealing algorithm on Hilbert space, and connected this to a diffusion process (SDE), a connection made in finite dimensions in [GH86]. The applications we are interested in are mainly on Bayesian nonparametrics; but of course, there are many more problems to which our techniques and results are applicable, *e.g.*, optimization, control theory, etc.
- The diffusion limits that we prove may not hold for discrete priors such as the Dirichlet pro-

cesses, or more generally Lévy random field priors [WCT11] because there is no underlying Gaussian measure. Likewise there is a growing interest for inverse problems in Besov priors based on wavelet bases and non-Gaussian random coefficients [SS⁺09]. Other Markov process limits are to be expected instead of a diffusion process. These class of models are very important in applications used in varied fields including computer science, machine learning and medical imaging. Thus it is of great importance to understand and quantify their efficiency; to achieve this will need new techniques and will be pursued elsewhere.

APPENDIX A: PROOFS OF LEMMAS; SECTION 4

PROOF OF LEMMA 5.1. Let us introduce the two 1-Lipschitz functions $h, h_* : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$h(x) = 1 \wedge e^x \quad \text{and} \quad h_*(x) = 1 + x 1_{\{x < 0\}}. \quad (\text{A.1})$$

The function h_* is a first order approximation of h in a neighbourhood of zero and we have

$$\alpha^\delta(x, \xi) = h\left(-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}\right) \quad \text{and} \quad \bar{\alpha}^\delta(x, \xi) = h_*\left(-\sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle\right)$$

where the proposal y is a function of x and ξ , as described in Equation (3.1). Since $h_*(\cdot)$ is close to $h(\cdot)$ in a neighbourhood of zero, the proof is finished once it is proved that $-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}$ is close to $-\sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle$. We have $\mathbb{E}_x[|\alpha^\delta(x, \xi) - \bar{\alpha}^\delta(x, \xi)|^p] \lesssim A_1 + A_2$ where the quantities A_1 and A_2 are given by

$$\begin{aligned} A_1 &= \mathbb{E}_x \left[\left| h\left(-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}\right) - h\left(-\sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle\right) \right|^p \right] \\ A_2 &= \mathbb{E}_x \left[\left| h\left(-\sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle\right) - h_*\left(-\sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle\right) \right|^p \right]. \end{aligned}$$

By Lemma 2.4, the first order Taylor approximation of Ψ is controled, $|\Psi(y) - \Psi(x) - \langle \nabla \Psi(x), y - x \rangle| \lesssim \|y - x\|_s^2$. The definition of the proposal y given in Equation (3.1) shows that $\|(y - x) - \sqrt{2\delta\tau}\xi\|_s \lesssim \delta\|x\|_s$. Assumptions 2.1 state that for $z \in \mathcal{H}^s$ we have $\langle \nabla \Psi(x), z \rangle \lesssim (1 + \|x\|_s) \cdot \|z\|_s$. Since the function $h(\cdot)$ is 1-Lipschitz it follows that

$$\begin{aligned} A_1 &= \mathbb{E}_x \left[\left| h\left(-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}\right) - h\left(-\sqrt{\frac{2\delta}{\tau}} \langle \nabla \Psi(x), \xi \rangle\right) \right|^p \right] \\ &\lesssim \mathbb{E}_x \left[\left| \Psi(y) - \Psi(x) - \langle \nabla \Psi(x), y - x \rangle \right|^p + \left| \langle \nabla \Psi(x), y - x - \sqrt{2\delta\tau}\xi \rangle \right|^p \right] \\ &\lesssim \mathbb{E}_x \left[\|y - x\|_s^{2p} + (1 + \|x\|_s^p) \cdot (\delta \|x\|_s)^p \right] \lesssim \delta^p (1 + \|x\|_s^{2p}). \end{aligned} \quad (\text{A.2})$$

Lemma 3.1 has been used to control the size of $\mathbb{E}_x[\|y - x\|^p]$. To bound A_2 , notice that for $z \in \mathbb{R}$ we have $|h(z) - h_*(z)| \leq \frac{1}{2} z^2$. Therefore the quantity A_2 can be bounded by

$$A_2 \lesssim \mathbb{E}_x \left[|\sqrt{\delta} \langle \nabla \Psi(x), \xi \rangle|^{2p} \right] \lesssim \delta^p \mathbb{E}_x \left[(1 + \|x\|_s^{2p}) \|\xi\|_s^{2p} \right] \lesssim \delta^p (1 + \|x\|_s^{2p}). \quad (\text{A.3})$$

Estimates (A.2) and (A.3) together give Equation (5.2). \square

PROOF OF COROLLARY 5.2. Let us prove Equations (5.3) and (5.4).

- Lemma 5.1 and Jensen's inequality give Equation (5.3).
- To prove (5.4), one can suppose $\delta^{\frac{p}{2}} \|x\|_s^p \leq 1$. Indeed, if $\delta^{\frac{p}{2}} \|x\|_s^p \geq 1$, we have

$$\mathbb{E}_x \left[|\alpha^\delta(x, \xi) - 1|^p \right] \lesssim 1 \leq \delta^{\frac{p}{2}} \|x\|_s^p \leq \delta^{\frac{p}{2}} (1 + \|x\|_s^p),$$

which gives the result. We thus suppose from now on that $\delta^{\frac{p}{2}} \|x\|_s \leq 1$. Under Assumptions 2.1 we have $\|\nabla \Psi(x)\|_{-s} \lesssim 1 + \|x\|_s$. Lemma 2.4 shows that for all $x, y \in \mathcal{H}^s$ we have $|\Psi(y) - \Psi(x) - \langle \nabla \Psi(x), y - x \rangle| \lesssim \|y - x\|_s^2$. The function $h(x) = 1 \wedge e^x$ is 1-Lipschitz, $\alpha^\delta(x, \xi) = h(-\frac{1}{\tau}[\Psi(y) - \Psi(x)])$ and $h(0) = 1$. Consequently,

$$\begin{aligned} \mathbb{E}_x \left[|\alpha^\delta(x, \xi) - 1|^p \right] &= \mathbb{E}_x \left[\left| h\left(-\frac{1}{\tau}[\Psi(y) - \Psi(x)]\right) - h(0) \right|^p \right] \\ &\lesssim \mathbb{E}_x \left[|\Psi(y) - \Psi(x)|^p \right] \lesssim \mathbb{E}_x \left[|\langle \nabla \Psi(x), y - x \rangle|^p + \|y - x\|_s^{2p} \right] \\ &\lesssim (1 + \|x\|_s^p) \cdot \mathbb{E}_x \left[\|y - x\|_s^p \right] + \mathbb{E}_x \left[\|y - x\|_s^{2p} \right]. \end{aligned}$$

By Lemma 3.1, for any integer $\beta \geq 1$ we have $\mathbb{E}_x \left[\|y - x\|_s^\beta \right] \lesssim \delta^\beta \|x\|_s^\beta + \delta^{\frac{\beta}{2}}$ so that the assumption $\delta^{\frac{p}{2}} \|x\|_s^p \leq 1$ leads to

$$\begin{aligned} \mathbb{E}_x \left[|\alpha^\delta(x) - 1|^p \right] &\lesssim (1 + \|x\|_s^p) \cdot (\delta^p \|x\|_s^p + \delta^{\frac{p}{2}}) + (\delta^{2p} \|x\|_s^{2p} + \delta^p) \\ &\lesssim (1 + \|x\|_s^p) \cdot (\delta^{\frac{p}{2}} + \delta^{\frac{p}{2}}) + (\delta^p + \delta^p) \\ &\lesssim \delta^{\frac{p}{2}} (1 + \|x\|_s^p). \end{aligned}$$

This finishes the proof of Corollary 5.2. □

PROOF OF LEMMA 5.4. The martingale difference $\Gamma^\delta(x, \xi)$ defined in Equation (3.8) can also be expressed as

$$\Gamma^\delta(x, \xi) = \xi + F(x, \xi)$$

where the error term $F(x, \xi) = F_1(x, \xi) + F_2(x, \xi)$ is given by

$$\begin{aligned} F_1(x, \xi) &= (2\tau\delta)^{-\frac{1}{2}} \left((1 - 2\delta)^{\frac{1}{2}} - 1 \right) (\gamma^\delta(x, \xi) - \mathbb{E}_x[\gamma^\delta(x, \xi)])x \\ F_2(x, \xi) &= (\gamma^\delta(x, \xi) - 1) \cdot \xi - \mathbb{E}_x[\gamma^\delta(x, \xi) \cdot \xi]. \end{aligned}$$

We now prove that the quantity $F(x, \xi)$ satisfies

$$\mathbb{E}_x \left[\|F(x, \xi)\|_s^2 \right] \lesssim \delta^{\frac{1}{4}} (1 + \|x\|_s^2) \quad (\text{A.4})$$

- We have $\delta^{-\frac{1}{2}} \left((1 - 2\delta)^{\frac{1}{2}} - 1 \right) \lesssim \delta^{\frac{1}{2}}$ and $|\gamma^\delta(x, \xi)| \leq 1$. Consequently,

$$\mathbb{E}_x \left[\|F_1(x, \xi)\|_s^2 \right] \lesssim \delta \|x\|_s^2 \quad (\text{A.5})$$

- Let us now prove that F_2 satisfies

$$\mathbb{E}_x \left[\|F_2(x, \xi)\|_s^2 \right] \lesssim \delta^{\frac{1}{4}} (1 + \|x\|^{\frac{1}{2}}). \quad (\text{A.6})$$

To this end, use the decomposition

$$\begin{aligned} \mathbb{E}_x \left[\|F_2(x, \xi)\|_s^2 \right] &\lesssim \mathbb{E}_x \left[|\gamma^\delta(x, \xi) - 1|^2 \cdot \|\xi\|_s^2 \right] + \|\mathbb{E}_x[\gamma^\delta(x, \xi) \cdot \xi]\|_s^2 \\ &= I_1 + I_2. \end{aligned}$$

Cauchy-Schwarz inequality shows that $I_1 \lesssim \mathbb{E}_x \left[|\gamma^\delta(x, \xi) - 1|^4 \right]^{\frac{1}{2}}$ where the Bernoulli random variable $\gamma^\delta(x, \xi)$ can be expressed as $\gamma^\delta(x, \xi) = \mathbb{I}_{\{U < \alpha^\delta(x, \xi)\}}$ where $U \stackrel{\mathcal{D}}{\sim} \text{Uniform}(0, 1)$ is independent from any other source of randomness. Consequently

$$\mathbb{E}_x \left[|\gamma^\delta(x, \xi) - 1|^4 \right] = \mathbb{E}_x \left[\mathbb{I}_{\{\gamma^\delta(x, \xi) = 0\}} \right] = 1 - \alpha^\delta(x)$$

where the mean local acceptance probability $\alpha^\delta(x)$ is defined by $\alpha^\delta(x) = \mathbb{E}_x[\alpha^\delta(x, \xi)] \in [0, 1]$. The convexity of the function $x \rightarrow |1 - x|$ ensures that

$$|1 - \alpha^\delta(x)| = |1 - \mathbb{E}_x[\alpha^\delta(x, \xi)]| \leq \mathbb{E}_x[|1 - \alpha^\delta(x, \xi)|] \lesssim \delta^{\frac{1}{2}} (1 + \|x\|)$$

where the last inequality follows from Corollary 5.2. This proves that $I_1 \lesssim \delta^{\frac{1}{4}} (1 + \|x\|^{\frac{1}{2}})$. To bound I_2 , it suffices to notice

$$\begin{aligned} I_2 &= \|\mathbb{E}_x[\gamma^\delta(x, \xi) \cdot \xi]\|_s^2 = \|\mathbb{E}_x[(\gamma^\delta(x, \xi) - 1) \cdot \xi]\|_s^2 \\ &\lesssim \mathbb{E}_x \left[|\gamma^\delta(x, \xi) - 1|^2 \cdot \|\xi\|_s^2 \right] = I_1 \end{aligned}$$

so that $I_2 \lesssim I_1 \lesssim \delta^{\frac{1}{4}} (1 + \|x\|^{\frac{1}{2}})$ and $\mathbb{E}_x \left[\|F_2(x, \xi)\|_s^2 \right] \lesssim \delta^{\frac{1}{4}} (1 + \|x\|^{\frac{1}{2}})$.

Combining Equation (A.5) and (A.6) gives Equation (A.4).

Let us now describe how Equations (5.7) and (5.8) follow from the estimate (A.4).

- We have $\mathbb{E}[\langle \hat{\varphi}_i, \xi \rangle_s \langle \hat{\varphi}_j, \xi \rangle_s] = \langle \hat{\varphi}_i, C_s \hat{\varphi}_j \rangle_s$ and $\mathbb{E}_x[\langle \hat{\varphi}_i, \Gamma^\delta(x, \xi) \rangle_s \langle \hat{\varphi}_j, \Gamma^\delta(x, \xi) \rangle_s] = \langle \hat{\varphi}_i, D^\delta(x) \hat{\varphi}_j \rangle_s$ with $\Gamma^\delta(x, \xi) = \xi + F(x, \xi)$. Consequently,

$$\begin{aligned} \langle \hat{\varphi}_i, D^\delta(x) \hat{\varphi}_j \rangle_s - \langle \hat{\varphi}_i, C_s \hat{\varphi}_j \rangle_s &= \mathbb{E}_x[\langle \hat{\varphi}_i, F(x, \xi) \rangle_s \langle \hat{\varphi}_j, F(x, \xi) \rangle_s] \\ &\quad + \mathbb{E}_x[\langle \hat{\varphi}_i, \xi \rangle_s \langle \hat{\varphi}_j, F(x, \xi) \rangle_s] \\ &\quad + \mathbb{E}_x[\langle \hat{\varphi}_i, F(x, \xi) \rangle_s \langle \hat{\varphi}_j, \xi \rangle_s]. \end{aligned}$$

We have $|\langle \hat{\varphi}_i, F(x, \xi) \rangle_s| \leq \|F(x, \xi)\|_s$ and Cauchy Schwarz's inequality proves that

$$\begin{aligned} \mathbb{E}_x[\langle \hat{\varphi}_i, F(x, \xi) \rangle_s \langle \hat{\varphi}_j, \xi \rangle_s]^2 &\leq \mathbb{E}_x[\|F(x, \xi)\|_s^2 \|\xi\|_s^2] \\ &\lesssim \mathbb{E}_x[\|F(x, \xi)\|_s^2]. \end{aligned}$$

It thus follows from Equation (A.4) that

$$\begin{aligned} |\langle \hat{\varphi}_i, D^\delta(x) \hat{\varphi}_j \rangle_s - \langle \hat{\varphi}_i, C_s \hat{\varphi}_j \rangle_s| &\lesssim \mathbb{E}_x[\|F(x, \xi)\|_s^2] + \mathbb{E}_x[\|F(x, \xi)\|_s^2]^{\frac{1}{2}} \\ &\lesssim \delta^{\frac{1}{8}} (1 + \|x\|_s), \end{aligned}$$

finishing the proof of (5.7).

- We have $\text{Trace}_{\mathcal{H}^s}(C_s) = \mathbb{E}[\|\xi\|_s^2]$ and $\text{Trace}_{\mathcal{H}^s}(D^\delta(x)) = \mathbb{E}[\|\Gamma^\delta(x, \xi)\|_s^2]$. Estimate (A.4) thus shows that

$$\begin{aligned}
|\text{Trace}_{\mathcal{H}^s}(D^\delta(x)) - \text{Trace}_{\mathcal{H}^s}(C_s)| &= |\mathbb{E}[\|\Gamma^\delta(x, \xi)\|_s^2 - \|\xi\|_s^2]| = |\mathbb{E}[\|\xi + F(x, \xi)\|_s^2 - \|\xi\|_s^2]| \\
&\lesssim |\mathbb{E}[\langle 2\xi + F(x, \xi), F(x, \xi) \rangle_s]| \lesssim \mathbb{E}[\|2\xi + F(x, \xi)\|_s \|F(x, \xi)\|_s] \\
&\lesssim \mathbb{E}[4\|\xi\|_s^2 + \|F(x, \xi)\|_s^2]^{\frac{1}{2}} \cdot \mathbb{E}[\|F(x, \xi)\|_s^2]^{\frac{1}{2}} \\
&\lesssim \left(1 + \delta^{\frac{1}{4}} (1 + \|x\|_s^2)\right)^{\frac{1}{2}} \cdot \left(\delta^{\frac{1}{8}} (1 + \|x\|_s)\right) \lesssim \delta^{\frac{1}{8}} (1 + \|x\|_s^2),
\end{aligned}$$

finishing the proof of (5.8). □

APPENDIX B: PROOF OF LEMMA 5.4

Before proceeding to give the proof, let us give a brief proof sketch. The proof of Lemma 6.4 consists in showing first that for any $\varepsilon > 0$ one can find a ball of radius $R(\varepsilon)$ around 0 in \mathcal{H}^s ,

$$B_0(R(\varepsilon)) = \{x \in \mathcal{H}_s : \|x\|_s \leq R(\varepsilon)\},$$

such that with probability $1 - 2\varepsilon$ we have $x^{k,\delta} \in B_0(R(\varepsilon))$ and $y^{k,\delta} \in B_0(R(\varepsilon))$ for all $0 \leq k \leq T\delta^{-1}$. As is described below, the existence of such a ball follows from the bound

$$\mathbb{E}[\sup_{t \in [0, T]} \|x(t)\|_s] < +\infty \tag{B.1}$$

where $t \mapsto x(t)$ is the solution of the stochastic differential equation (3.6). For the sake of completeness, we include a proof of Equation (B.1). The solution $t \mapsto x(t)$ of the stochastic differential equation (3.6) satisfies $x(t) = \int_0^t d(x(u)) du + \sqrt{2\tau} W(t)$ for all $t \in [0, T]$ where the drift function $d(x) = -(x + C\nabla\Psi(x))$ is globally Lipschitz on \mathcal{H}^s , as described in Lemma 2.4. Consequently $\|d(x)\|_s \leq A(1 + \|x\|_s)$ for some positive constant $A > 0$. The triangle inequality then shows that

$$(B.2) \quad \|x(t)\|_s \leq A \int_0^t (1 + \|x(u)\|_s) du + \sqrt{2\tau} \|W(t)\|_s.$$

By Gronwall's inequality we obtain

$$(B.3) \quad \sup_{[0, T]} \|x(t)\|_s \leq (AT + \sup_{[0, T]} \|W(t)\|_s) [1 + ATe^{AT}].$$

Since $\mathbb{E}[\sup_{[0, T]} \|W(t)\|_s] < \infty$, the bound (B.1) is proved.

PROOF OF LEMMA 6.4. The proof consists in showing that the acceptance probability of the algorithm is sufficiently close to 1 so that approximation $t^\delta(k) \approx k$ holds. The argument can be divided into 3 main steps. In the first part, we show that we can find a finite ball $B(0, R(\varepsilon))$ such that the trajectory of the Markov chain $\{x^{k,\delta}\}_{k \leq T\delta^{-1}}$ remains in this ball with probability at least $1 - 2\varepsilon$. This observation is useful since the function Ψ is Lipschitz on any ball of finite radius in \mathcal{H}^s . In the second part, using the fact that Ψ is Lipschitz on $B(0, R(\varepsilon))$, we find a lower bound for the acceptance probability α^δ . Then, in the last step, we use a moment estimate to prove that one can make the lower bound uniform on the interval $0 \leq k \leq T\delta^{-1}$.

- **Restriction to a Ball of Finite Radius**

First, we show that with high probability the trajectory of the MCMC algorithm stays in a ball of finite radius. The functional $x \mapsto \sup_{t \in [0, T]} \|x(t)\|_s$ is continuous on $C([0, T], \mathcal{H}_s)$ and $\mathbb{E}[\sup_{t \in [0, T]} \|x(t)\|_s] < \infty$ for $t \mapsto x(t)$ following the stochastic differential equation (3.6), as proved in Equation (B.1). Consequently, the weak convergence of z^δ to the solution of (3.6) encapsulated in Theorem 3.2 shows that $\mathbb{E}[\sup_{k < T\delta^{-1}} \|x^{k, \delta}\|_s]$ can be bounded by a finite universal constant independent from δ . Given $\varepsilon > 0$, Markov inequality thus shows that one can find a radius $R_1 = R_1(\varepsilon)$ large enough so that the inequality

$$\mathbb{P}[\|x^{k, \delta}\|_s < R_1 \quad \text{for all} \quad 0 \leq k \leq T\delta^{-1}] > 1 - \varepsilon \quad (\text{B.4})$$

for any $\delta \in (0, \frac{1}{2})$. By Fernique's Theorem there exists $\alpha > 0$ such that $\mathbb{E}[e^{\alpha \|\xi\|_s^2}] < \infty$. This implies that $\mathbb{P}[\|\xi\|_s > r] \lesssim e^{-\alpha r^2}$. Therefore, if $\{\xi_k\}_{k \geq 0}$ are i.i.d. Gaussian random variables distributed as $\xi \stackrel{\mathcal{D}}{\sim} \pi_0$, the union bound shows that

$$\mathbb{P}[\|\sqrt{\delta}\xi_k\|_s \leq r \quad \text{for all} \quad 0 \leq k \leq T\delta^{-1}] \gtrsim 1 - T\delta^{-1} \exp(-\alpha \delta^{-1} r^2).$$

This proves that one can choose $R_2 = R_2(\varepsilon)$ large enough in such a manner that

$$\mathbb{P}[\|\sqrt{\delta}\xi_k\|_s < R_2 \quad \text{for all} \quad 0 \leq k \leq T\delta^{-1}] > 1 - \varepsilon \quad (\text{B.5})$$

for any $\delta \in (0, \frac{1}{2})$. At temperature $\tau > 0$ the MCMC proposals are given by $y^{k, \delta} = (1 - 2\delta)^{\frac{1}{2}} x^{k, \delta} + (2\delta\tau)^{\frac{1}{2}} \xi_k$. It thus follows from the bounds (B.4) and (B.5) that with probability at least $(1 - 2\varepsilon)$ the vectors $x^{k, \delta}$ and $y^{k, \delta}$ belong to the ball $B_0(R(\varepsilon)) = \{x \in \mathcal{H}_s : \|x\|_s < R(\varepsilon)\}$ for $0 \leq k \leq T\delta^{-1}$ where radius $R(\varepsilon)$ is given by $R(\varepsilon) = R_1(\varepsilon) + R_2(\varepsilon)$.

- **Lower Bound for Acceptance Probability**

We now give a lower bound for the acceptance probability $\alpha^\delta(x^{k, \delta}, \xi^k)$ that the move $x^{k, \delta} \rightarrow y^{k, \delta}$ is accepted. Assumptions 2.1 state that $\|\nabla \Psi(x)\|_{-s} \lesssim 1 + \|x\|_s$. Therefore, the function $\Psi : \mathcal{H}^s \rightarrow \mathbb{R}$ is Lipschitz on $B_0(R(\varepsilon))$,

$$\|\Psi\|_{\text{lip}, \varepsilon} \stackrel{\text{def}}{=} \sup \left\{ \frac{|\Psi(y) - \Psi(x)|}{\|y - x\|_s} : x, y \in B_0(R(\varepsilon)) \right\} < \infty.$$

One can thus bound the acceptance probability $\alpha^\delta(x^{k, \delta}, \xi^k) = 1 \wedge \exp(-\tau^{-1}[\Psi(y^{k, \delta}) - \Psi(x^{k, \delta})])$ for $x^{k, \delta}, y^{k, \delta} \in B_0(R(\varepsilon))$. Since the function $z \mapsto 1 \wedge e^{-\tau^{-1}z}$ is Lipschitz with constant τ^{-1} , the definition of $\|\Psi\|_{\text{lip}, \varepsilon}$ shows that the bound

$$\begin{aligned} 1 - \alpha^\delta(x^{k, \delta}, \xi^k) &\leq \tau^{-1} \|\Psi\|_{\text{lip}, \varepsilon} \|y^{k, \delta} - x^{k, \delta}\|_s \\ &\leq \tau^{-1} \|\Psi\|_{\text{lip}, \varepsilon} \left\{ [(1 - 2\delta)^{\frac{1}{2}} - 1] \|x^{k, \delta}\|_s + (2\delta\tau)^{\frac{1}{2}} \|\xi^k\|_s \right\} \\ &\lesssim \sqrt{\delta} (1 + \|\xi^k\|_s) \end{aligned}$$

holds for every $x^{k, \delta}, y^{k, \delta} \in B_0(R(\varepsilon))$. Hence, there exists a constant $K = K(\varepsilon)$ such that $\hat{\alpha}^\delta(\xi^k) = 1 - K\sqrt{\delta} (1 + \|\xi^k\|_s)$ satisfies $\alpha^\delta(x^{k, \delta}, \xi^k) > \hat{\alpha}^\delta(\xi^k)$ for every $x^{k, \delta}, y^{k, \delta} \in B_0(R(\varepsilon))$. Since the trajectory of the MCMC algorithm stays in the ball $B_0(R(\varepsilon))$ with probability at least $1 - 2\varepsilon$ the inequality

$$\mathbb{P}[\alpha^\delta(x^{k, \delta}, \xi^k) > \hat{\alpha}^\delta(\xi^k) \quad \text{for all} \quad 0 \leq k \leq T\delta^{-1}] > 1 - 2\varepsilon.$$

holds for every $\delta \in (0, \frac{1}{2})$.

- **Second Moment Method**

To prove that $t^\delta(k)$ does not deviate too much from k , we show that its expectation satisfies $\mathbb{E}[t^\delta(k)] \approx k$ and we then control the error by bounding the variance. Since the Bernoulli random variable $\gamma^{k,\delta} = \text{Bernoulli}(\alpha^\delta(x^{k,\delta}\xi^k))$ are not independent, the variance of $t^\delta(k) = \sum_{l \leq k} \gamma^{l,\delta}$ is not easily computable. We thus introduce i.i.d. auxiliary random variables $\hat{\gamma}^{k,\delta}$ such that

$$\sum_{l \leq k} \hat{\gamma}^{l,\delta} = \hat{t}^\delta(k) \approx t^\delta(k) = \sum_{l \leq k} \gamma^{l,\delta}.$$

As described below, the behaviour of $\hat{t}^\delta(k)$ is readily controlled since it is a sum of i.i.d. random variables. The proof then exploits the fact that $\hat{t}^\delta(k)$ is a good approximation of $t^\delta(k)$.

The Bernoulli random variables $\gamma^{k,\delta}$ can be described as $\gamma^{k,\delta} = \mathbb{I}(U_k < \alpha^\delta(x^{k,\delta}\xi^k))$ where $\{U_k\}_{k \geq 0}$ are i.i.d. random variables uniformly distributed on $(0, 1)$. As a consequence, with probability at least $1 - 2\varepsilon$, the random variables $\hat{\gamma}^{k,\delta} = \mathbb{I}(U_k < \hat{\alpha}^\delta)$ satisfy $\gamma^{k,\delta} \geq \hat{\gamma}^{k,\delta}$ for all $0 \leq k \leq T\delta^{-1}$. Therefore, with probability at least $1 - 2\varepsilon$, we have $t^\delta(k) \geq \hat{t}^\delta(k)$ for all $0 \leq k \leq T\delta^{-1}$ where $\hat{t}^\delta(k) = \sum_{l \leq k} \hat{\gamma}^{l,\delta}$. Consequently, since $t^\delta(k) \leq k$, to prove Lemma 6.4 it suffices to show instead that the following limit in probability holds,

$$\lim_{\delta \rightarrow 0} \sup \{ \delta \cdot |\hat{t}^\delta(k) - k| : 0 \leq k \leq T\delta^{-1} \} = 0. \quad (\text{B.6})$$

Contrary to the random variables $\{\gamma^{k,\delta}\}_{k \geq 0}$, the random variables $\{\hat{\gamma}^{k,\delta}\}_{k \geq 0}$ are i.i.d. and are thus easily controlled. By Doob's inequality we have

$$\mathbb{P} \left[\sup \{ \delta \cdot |\hat{t}^\delta(k) - \mathbb{E}[\hat{t}^\delta(k)]| : 0 \leq k \leq T\delta^{-1} \} > \eta \right] \leq 2 \frac{\text{Var}(\hat{t}^\delta(T\delta^{-1}))}{(\delta^{-1}\eta)^2} \leq 2 \frac{\delta T}{\eta^2}.$$

Since $\mathbb{E}[\hat{t}^\delta(k)] = k \cdot \{1 - K\sqrt{\delta} (1 + \mathbb{E}[\|\xi^k\|_s])\}$, Equation (B.6) follows. This finishes the proof of Lemma 6.4. □

Acknowledgements. The authors are grateful to David Dunson for the his comments on the implications of theory, Frank Pinski for helpful discussions concerning the behaviour of the quadratic variation; these discussions crystalized the need to prove Theorem 6.3. The authors are also grateful to the three anonymous referees, an associate editor and the editor for their insightful comments which helped us improve the exposition. NSP gratefully acknowledges the NSF grant DMS 1107070. AMS is grateful to EPSRC and ERC for financial support. Parts of this work was done when AHT was visiting the department of Statistics at Harvard university. The authors thank the department of statistics, Harvard University for its hospitality.

REFERENCES

- [Ber86] E. Berger. Asymptotic behaviour of a class of stochastic approximation procedures. *Probab. Theory Relat. Fields*, 71(4):517–552, 1986.
- [BKMS08] E. Buckwar, R. Kuske, S.E. Mohammed, and T. Shardlow. Weak convergence of the euler scheme for stochastic differential delay equations. *LMS Journal of Computation and Mathematics*, 11(-1):60–99, 2008.
- [BRSV08] A. Beskos, G.O. Roberts, A.M. Stuart, and J. Voss. An MCMC method for diffusion bridges. *Stochastics and Dynamics*, 8(3):319–350, 2008.

- [BS05] E. Buckwar and T. Shardlow. Weak approximation of stochastic differential delay equations. *IMA journal of numerical analysis*, 25(1):57, 2005.
- [BT93] D. Bertsimas and J. Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.
- [Cas08] I. Castillo. Lower bounds for posterior rates with gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.
- [CDS11] SL Cotter, M. Dashti, and AM Stuart. Variational data assimilation using targetted random walks. *International Journal for Numerical Methods in Fluids*, 2011.
- [Čer85] V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.
- [CH06] A.J. Chorin and O.H. Hald. *Stochastic Tools in Mathematics and Science*. Springer Verlag, 2006.
- [CRSW11] S.L. Cotter, G.O. Roberts, A.M. Stuart, and D. White. Mcmc methods for functions: modifying old algorithms to make them run faster. *Preprint*, 2011.
- [Dac89] B. Dacorogna. *Direct Methods in the Calculus of Variations*. Springer, New York, 1989.
- [dB06] A. de Bouard and A. Debussche. Weak and strong order of convergence of a semidiscrete scheme for the stochastic nonlinear schrodinger equation. *Applied Mathematics and Optimization*, 54(3):369–399, 2006.
- [DP09] A. Debussche and J. Printems. Weak order for the discretization of the stochastic heat equation. *Math. Comp*, 78(266):845–863, 2009.
- [DPP07] D.B. Dunson, N.S. Pillai, and J.H. Park. Bayesian density regression”. *Journal of the Royal Statistical Society, Series-B*, 69:163–183, 2007.
- [DPZ92] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.
- [DPZ96] G. Da Prato and J. Zabczyk. *Ergodicity for Infinite Dimensional Systems*. Cambridge Univ Pr, 1996.
- [DS11] M. Dashti and A.M. Stuart. Title. *In preparation*, 2011.
- [EHN96] H.K. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.
- [Fit91] B.G. Fitzpatrick. Bayesian analysis in inverse problems. *Inverse problems*, 7:675, 1991.
- [Gem85] D. Geman. Bayesian image analysis by adaptive annealing. *IEEE Transactions on Geoscience and Remote Sensing*, 1:269–276, 1985.
- [GH86] S. Geman and C.R. Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24:1031, 1986.
- [GKL09] M. Geissert, M. Kovács, and S. Larsson. Rate of weak convergence of the finite element method for the stochastic heat equation with additive noise. *BIT Numerical Mathematics*, 49(2):343–356, 2009.
- [GVDV11] S. Ghosal and A. Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2011.
- [Hau10] E. Hausenblas. Weak approximation of the stochastic wave equation. *Journal of computational and applied mathematics*, 235(1):33–58, 2010.
- [HAVW05] M. Hairer, A.M. Stuart, J. Voss, and P. Wiberg. Analysis of SPDEs arising in path sampling. Part 1: the gaussian case. *Comm. Math. Sci.*, 3:587–603, 2005.
- [Hen81] D. Henry. *Geometric Theory of Semilinear Parabolic Equations*, volume 61. Springer-Verlag, 1981.
- [HKS89] R.A. Holley, S. Kusuoka, and D.W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *Journal of Functional Analysis*, 83(2):333–347, 1989.
- [HPUU08] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*. Springer Verlag, 2008.
- [HSV07a] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling, part II: The nonlinear case. *Annals of Applied Probability*, 17:1657–1706, 2007.
- [HSV07b] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling. PartII: the nonlinear case. *Ann. Appl. Probab.*, 17(5-6):1657–1706, 2007.
- [HSV10] M. Hairer, A. M. Stuart, and J. Voss. Signal processing problems on function space: Bayesian formulation, stochastic pdes and effective mcmc methods. *The Oxford Handbook of Nonlinear Filtering*, Editors D. Crisan and B. Rozovsky, 2010. To Appear.
- [HSV11] M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for metropolis-hastings algorithms in infinite dimensions. 2011. <http://arxiv.org/abs/1112.1392>.
- [IZ00] H. Ishwaran and M. Zarepour. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87:371–390, 2000.
- [KJV83] S. Kirkpatrick, D.G. Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [KLL10] M. Kovács, S. Larsson, and F. Lindgren. Weak convergence of finite element approximations of linear

- stochastic evolution equations with additive noise. *preprint*, 2010.
- [MPS11] J.C. Mattingly, N.S. Pillai, and A.M. Stuart. SPDE Limits of the Random Walk Metropolis Algorithm in High Dimensions. *Ann. Appl. Prob.*, 2011.
 - [MT93] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.
 - [MW04] J. Møller and R.P. Waagepetersen. *Statistical inference and simulation for spatial point processes*, volume 100. CRC Press, 2004.
 - [PR08] O. Papaspiliopoulos and G.O. Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.
 - [RC04] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.
 - [RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.
 - [RR98] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998.
 - [RR01] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.
 - [RS01] G.O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the metropolis-hastings algorithm. *Biometrika*, 88(3):603–621, 2001.
 - [RW06] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
 - [Sha03] T. Shardlow. Weak convergence of a numerical method for a stochastic heat equation. *BIT Numerical Mathematics*, 43(1):179–193, 2003.
 - [SS⁺09] M.L. Saksman, S. Siltanen, et al. Discretization-invariant bayesian inversion and besov space priors. *Inverse Problems and Imaging*, 3:87–122, 2009.
 - [Stu10] AM Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19(-1):451–559, 2010.
 - [Tie98] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.
 - [VDVVZ08] A. Van Der Vaart and J.H. Van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *Annals of Statistics*, 36(3):1435–1463, 2008.
 - [WCT11] Robert L. Wolpert, Merlise A. Clyde, and Chong Tu. Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. *Annals of Statistics*, page (in press), 2011.
 - [Zha00] L.H. Zhao. Bayesian aspects of some nonparametric problems . *Annals of Statistics*, 28(2):532–552, 2000.

DEPARTMENT OF STATISTICS,
HARVARD UNIVERSITY
1 OXFORD STREET, CAMBRIDGE
02138, MA, USA
E-MAIL: pillai@stat.harvard.edu

MATHEMATICS INSTITUTE
WARWICK UNIVERSITY
CV4 7AL, UK
E-MAIL: a.m.stuart@warwick.ac.uk

DEPARTMENT OF STATISTICS
WARWICK UNIVERSITY
CV4 7AL, UK
E-MAIL: a.h.thiery@warwick.ac.uk